jamk.fi

Big data
Ratkaisut ja testaaminen

Jero Marjalahti

Opinnäytetyö
Elokuu 2016
Tekniikan ja liikenteen ala
Insinööri (AMK), tietotekniikka

Jyväskylän ammattikorkeakoulu
JAMK University of Applied Sciences

jamk.fi

Kuvailulehti

Tekijä(t)	Julkaisun laji	Päivämäärä
Marjalahti, Jero	Opinnäytetyö, AMK	21.08.2016
	Sivumäärä 105	Julkaisun kieli Suomi
		Verkkojulkaisulupa myönnetty: x

Työn nimi
Big data – Ratkaisut ja testaaminen

Tutkinto-ohjelma
Insinööri (AMK), tietotekniikka

Työn ohjaaja(t)
Mika Rantonen, Antti Häkkinen

Toimeksiantaja(t)
JYVSECTEC, Marko Vatanen

Tiivistelmä

Opinnäytetyön tehtävänä oli JYVSECTEC:in toimesta tutustua yleisellä tasolla big dataan sekä tutkia tarkemmin viiden suurimman big data -toimittajan ratkaisuja ja testata opinnäytetyön edistyessä valittuja ratkaisuja.

Big datan teoriaosassa käytiin läpi big datan käytössä olevat määritelmät, historia sekä siihen johtaneet muutokset datan määrän kasvussa ja monimuotoisuudessa. Big datan arkkitehtuurista selvitettiin siihen liittyvät rajapinnat ja datalähteet. Eri palvelutyypeissä tutustuttiin big data -palveluiden organisointiin ja työkaluihin sekä ICT-alan suuryritysten luomiin ja käyttöönottamiin ratkaisuihin.

Etukäteen valittujen viiden suuriman big data -toimittajan jakeluista käytiin läpi teoriatasolla toimittajien heikkoudet ja vahvuudet, tarjolla olevat eri vaihtoehdot sekä niiden ohjelmisto- ja laitteistovaatimukset.

Korkeiden kustannusten ja suurten laitteistovaatimusten takia testauksessa päädyttiin tutkimaan tarkemmin IBM:n Bluemix-pilvipalvelun tarjoamia mahdollisuuksia. Tätä päätöstä tuki myös se, että IBM:ltä saatiin opiskelijatunnukset, jotka mahdollistivat pidempiaikaisen testauksen. Testauksessa käytettiin Streaming Analytics, Insights for Twitter ja Insights for Weather -palveluja, joilla saatiin big dataan liittyvät testisovellukset luotua ja testattua.

Avainsanat (asiasanat)
Big data, Hadoop, tiedonhallinta, tilastotiede, matemaattinen tilastotiede, analyysi, liiketoimintatiedon hallinta

Muut tiedot

jamk.fi

Description

Author(s)	Type of publication	Date
Marjalahti, Jero	Bachelor's thesis	21.08.2016
		Language of publication: Finnish
	Number of pages 105	Permission for web publication: x

Title of publication
Big Data – Solutions and Testing

Degree programme
Information Technology

Supervisor(s)
Rantonen Mika, Häkkinen Antti

Assigned by
JYVSECTEC, Vatanen Marko

Abstract

The thesis was assigned by JYVSECTEC and it focused on Big Data on a general level. The study researched and tested the solutions of the five biggest big data companies.

The process started with the basic theory, common terminology, history and the changes of data occurring now in Big Data. The architecture part of Big Data reviews the interfaces and data sources, and it is followed by different kinds of Big Data management solutions, tools and the major ICT enterprises' own Big Data solutions.

The five major Big Data solutions selected in advance were processed in theory including their weaknesses and strengths, different kind of options they offer as well as the software and hardware requirements.

Because of the high expense and the high hardware requirements the testing part ended up with a detailed research of IBM's Bluemix cloud solution and its options. This decision was supported by the academic code which was given by the IBM. With the academic code it was possible to test Bluemix for a longer period of time. The services used in the testing part were Streaming Analytics, Insights for Twitter and Insights for Weather. With these services it was possible to create and test Big Data applications.

Keywords/tags (subjects)
Big data, Hadoop, information management, statistics, mathematical statistics, analysis, business intelligence

Miscellaneous

Sisältö

Kuviot

Taulukot

Lyhenteet

AMI	Amazon Machine Image
AMPP	Asymmetric Massively Parallel Processing
ANSI	American National Standards Institute
API	Application Programming Interfaces
AWS	Amazon Web Services
CDH	Cloudera Distribution Including Apache Hadoop
CF CLI	Cloud Foundry Command-Line Interface
CLI	Command-Line Interface
DIA	Data Integration Appliance
DSL	Domain-Specific Languages
EDH	Enterprise Data Hub
EDW	Enterprise Data Warehouse
FPGA	Field-Programmable Gate Array
GCP	Google Cloud Platform
GP HD	Greenplum Hadoop
GP HDEE	Greenplum Hadoop Enterprise Edition
GP MR	Greenplum Hadoop MapR
HA/DR	High Availability/Disaster Recovery
HDFS	Hadoop Distributed File System
HDP	Hortonworks Data Platform
HPLI	Hadoop Physical Infrastructure Layer
IaaS	Infrasturcture as a Service
IOPS	Input/Output Operations Per Second

JACL	Java Command Language
JSON	JavaScript Object Notation
KMS	Key Management Sevice
LDAP	Lightweight Directory Access Protocol
MPP	Massively Parallel Processing
NFS	Network File System
NoSQL	Not only Structured Query Language
ODPi	Open Data Platform intiative
PaaS	Platform as a Service
PCF	Pivotal Cloud Foundry
PCIe	Peripheral Component Interconnect express
POSIX	Portable Operating System Interface for uniX
RAM	Random Access Memory
RDBMS	Relational Database Management System
RDS	Relational Database Sevice
REST	REpresentational State Transfer
RGCE	Realistic Global Cyber Environment
RHEL	Red Hat Enterprise Linux
ROM	Read Only Memory
RPM	Red Hat Package Manager
SaaS	Software as a Sevice
SLA	Service-Level Agreement
SQL	Structured Query Language
SSD	Solid-State Drive

SSL	Secure Sockets Layer
TLS	Transport Layer Security
UAP	Unified Analytics Platform
VM	Virtual Machine
VMDK	Virtual Machine Disk
VPN	Virtual Private Network
XML	Extensible Markup Language

1 Työn kuvaus

1.1 Toimeksiantajan esittely

Toimeksiantajana opinnäytetyölle toimi riippumaton kyberturvallisuuden tutkimus-, kehitys- ja koulutuskeskus JYVSECTEC - Jyväskylä Security Technology. JYVSECTEC on perustettu projektimuotoisena vuonna 2011 Jyväskylän ammattikorkeakoulun IT-instituutissa ja se keskittyy kyberturvallisuusharjoituksiin sekä tarjoaa konsultointi-, tutkimus-, testaus- ja koulutuspalveluja. Lisäksi JYVSECTEC toteuttaa tutkimus- ja kehityshankkeita. Yrityksellä on käytössä realistisesti mallinnettu RGCE (Realistic Global Cyber Environment) kybertoimintaympäristö. JYVSECTECin päämääränä on luoda Keski-Suomeen yksi Suomen johtavista kyberturvallisuuden tutkimus-, kehitys-, ja koulutuskeskuksista sekä kehittää turvallisuusalan kansallista ja kansainvälistä yritysten ja toimijoiden yhteistyöverkostoa. (Jyvsectec - Tietoa meistä 2016)

1.2 Työn tavoitteet

Opinnäytetyön tavoitteena oli tutustua yleisellä tasolla big dataan sekä tutkia tarkemmin tarjolla olevien big dataan liittyvien eri toimittajien ratkaisuja. Opinnäytetyötä varten valittiin aluvasti viisi eri palveluntarjoajaa ja heidän tuotteensa:

- Cloudera Enterprise - Cloudera Distribution Including Apache Hadoop
- Hortonworks - Hortonworks Data Platform
- IBM - BigInsights for Apache Hadoop
- MapR - MapR Converged Data Platform
- Pivotal - Pivotal Big Data Suite

Näistä etukäteen sovituista tuotteista valikoidaan sopivat jakelut tarkempaa testausta varten.

2 Mitä on big data?

Arvion mukaan maailmassa on tällä hetkellä tallennettua dataa noin 5-10 tsettatavua (ZB). Datan määrän arvioidaan kasvavan 2,5 eksatavua (EB) päivittäin ja vuonna 2020 datan määrä olisi jo 40 tsettatavua. Arvion mukaan 90 % tämän hetkisestä datasta olisi luotu viimeisen kahden vuoden aikana. Suurimmat datan tallentajat ovat Google, Microsoft, Facebook sekä Amazon. Valtavalla vauhdilla kasvava datan tallentaminen nyky-yhteiskunnassa on luonut ilmiön nimeltä big data. (Cloudtweaks - Surprising Facts and Stats About The Big Data Industry 2015)

Käsitteenä big data on noussut esille vuoden 2005 paikkeilla, mutta todellinen läpimurto big datalle tapahtui vuonna 2011. Pelkästään valtava määrä dataa ei yksinään ole big dataa. Big data tarkoittaa valtavia tietomääriä sisältävää tallennettua dataa, joka lisääntyy nopeasti ja on monimuotoista. Big data ei ole käsitteenä kuitenkaan täysin selkeä. Siitä on vaikea tehdä yhtenevää määritelmää, joka sopisi kaikkiin tilanteisiin, kun dataa on tallennettuna paljon. Yleisesti puhutaan big datasta, kun nopeasti lisääntyvää loogisesti järjestelemätöntä tietoa on paljon. Näin ollen tiedon analysointi, hakeminen, käyttö ja hyödyntäminen tehokkaasti on hankalaa sekä hidasta perinteisellä tiedonhallintateknologialla. Kaikkea dataa ei kuitenkaan tarvitse saada analysoitua, vaan vain tärkeä murto-osa, josta saadaan riittävästi informaatiota muodostamaan tarvittava tieto. Näin ollen tilastomatematiikka onkin oleellinen osa big data -ratkaisuja. Big datasta puhuessa sillä voidaan myös tarkoittaa siihen liittyviä tuotteita, palveluja ja tekniikoita. (Salo 2014, 6-10, 26, 31-32; Bigdata - Big data -määritelmiä n.d.)

Suuri haaste big datassa tänä päivänä on reagointi reaaliaikaisesti analysoituun dataan sekä sen nopea hyödyntäminen käytännössä. Data voi olla yksittäisiä suuria tiedostoja kuten korkealaatuista audiovisuaalista materiaalia tai data voi olla myös kooltaan hyvin pieniäkin tiedostoja. Esimerkiksi sääyhtiölle hetkellisistä säätiedoista ja loppukäyttäjän sääsovelluksesta koostuvasta pienistä tiedoista kertyy reaaliaikaisena toteutuksena dataa useamman teratavun (TB) verran yhden tunnin aikana. (Forbes - How Real-Time Weather Data Is Helping Businesses Run Better 2015)

Big data -käsitteen hahmottamista helpottaa seuraavanlainen Bigdata.fi -sivustolta löytyvä lista:

- **Volume**
  Dataa on paljon.
- **Velocity**
  Dataa tulee vauhdilla lisää ja päätöksiä pitäisi tehdä nopeasti.
- **Variety**
  Data on yhä monimuotoisempaa ("85 % datasta on strukturoimatonta eli vailla selkeää rakennetta").
- **Value**
  Arvottavasta dataheinäsuovasta pitäisi löytää ymmärryksen neula.
- **Veracity**
  Onko data mielekästä tai arvokasta käsiteltävän ongelman kannalta.
- **Volatility**
  Kauanko data on olennaista ja miten pitkään sitä pitäisi säilyttää.

Tästä listasta yleisimmäksi big datalle nousee kolmen v-kirjaimen määritelmä volume, velocity ja variety eli suomeksi käännettynä volyymi, vauhti ja vaihtelevuus. (Bigdata - Big Data -määritelmiä n.d.)

Big data on tällä hetkellä yksi tärkeimmistä teknologiatrendeistä. Sillä on potentiaalia dramaattisesti muuttaa yritysten ja organisaatioiden tapa hyödyntää tallentuva data liiketoimintamalleihin. Olemassa olevasta datasta tarjoutuu mahdollisuus tutkia yhtäläisyyksiä ja kuvioita. Tämä taas tarjoaa yrityksille aivan uudenlaisen tavan vastata jo aikaisessa vaiheessa esimerkiksi asiakastyytyväisyyteen ja ostotottumuksien muuttumiseen. Teollisuudessa tuotannon puolella datan analysointi laitteistoiden sensoreista mahdollistaa huomaamaan ongelmat ajoissa ja näin ollen estämään ne ennen isompia ongelmia tai jopa tuotannon kallista keskeytymistä. (Hurwitz, Nugent, Halper & Kaufman 2013, Introduction)

Niin isot kuin pienetkin yritykset kaikilla toimialoilla ovat aina saaneet suuren hyödyn datanhallinnasta. Yritysten ja niiden toimialojen kasvaessa ja monipuolistuessa asiakkaista, tuotteista sekä palveluista kerättävä tarvittava informaatio on kuitenkin tuottanut vaikeuksia yrityksille. Samat ongelmat koskevat myös tutkimus- ja kehityspuolta, kuten esimerkiksi laskentatehon puute monimutkaisten mallien kohdalla tai kuvien prosessoinnissa dataksi. (Hurwitz ym. 2013, Grasping the Fundamentals of Big Data)

Osa datasta strukturoidaan eli järjestellään vaadittavien ehtojen mukaisesti, mutta suurin osa, mukaan lukien dokumentit, asiakaspalvelutiedot ja jopa kuvat sekä videot, ovat strukturoimatonta. On myös aivan uudenlaista dataa, kuten sosiaalisesta mediasta ja verkkosivujen lokitiedostoista tulevaa. Kun monimuotoista dataa tulee valtavat määrät vauhdilla, on mahdotonta käyttää pelkästään perinteisiä tiedonhallintamenetelmiä, mikäli halutaan hyötyä tästä mahdollisuudesta. Big datan mahdollisuus, mutta samalla myös haaste, on hallita data eri lailla kuin ennen perinteisillä työkaluilla. Mikäli yritykset kykenevät analysoimaan petatavuja (PB) dataa hyväksyttävällä sekä kustannustehokkaalla tavalla erottaakseen datasta malleja tai poikkeavuuksia, se luo täysin uusia mahdollisuuksia hyödyntää dataa tietona. (Hurwitz ym. 2013, Understanding the Waves of Managing Data)

Kaikesta huolimatta pitää myös ymmärtää, että big datan analysointi ei ole pelkästään yrityksien uusi mahdollisuus. Esimerkkeinä mainittakoon, että lääketiede, tähtitiede ja terrorismin vastainen työ keräävät myös tällä hetkellä käsittämättömiä määriä dataa. Big datan analysoinnilla voidaan siis myös ennen kaikkea pelastaa ihmishenkiä taloudellisen hyödyn lisäksi. (Mt.)

Dataa pitää myös osata lähestyä eri tavoin, mikäli se on hetkessä muuttuvaa tai levossa olevaa. Muuttuva data voi olla esimerkiksi yrityksen tuotteen laadun reaaliaikaista valvontaa, ja näin ollen mahdollistetaan puuttuminen laatupoikkeamiin ennen taloudellista takaiskua yritykselle. Levossa oleva data voi olla esimerkiksi yrityksen tämän hetkiset kuluttajien ostokuviot mukaan lukien ostotottumukset, sosiaalinen media ja asiakastyytyväisyyskyselyt. (Mt.)

Big dataan ei ole olemassa yhtä oikeaa sovellusta tai työkalua. Kaikki yhteen sitoutuvat päällekkäiset teknologiat yhdessä antavat oikeaan tietoon perustuvan ymmärryksen oikeana hetkenä, oli se sitten ihmisten, koneiden tai internetin luomaa dataa. Tässä on kuitenkin huomioitava datan todenmukaisuus ja arvo. Onko analysoinnin tarkkuus riittävä sen hyödyntämisen tuottamalle arvolle ja onko analyysin tuloksessa lopulta mitään järkeä. (Hurwitz ym. 2013, Defining Big Data)

Big datassa tärkeä näkökulma on siis yhteiskunnallinen sekä taloudellinen hyöty, joka usein jää huomioimatta teknologian varjosta. Yrityksille big data luo uusia liiketoimintamahdollisuuksia, mutta riskit epäonnistumiseen big datan suhteen ovat suuret. Big dataa ei pitäisi käsitellä erillisenä osana yrityksissä, vaan se pitäisi sisällyttää osana liiketoiminta-analytiikkaan. Vaikka yksi big dataan liittyvistä väitteistä onkin, että yritystoiminta tulee kärsimään kehityksessä ilman big datan hyödyntämistä, sille pitäisi asettaa kuitenkin samat realistiset hyöty- ja tuotto-odotukset kuin muullekin yrityksen analytiikalle. (Salo 2014, 38)

3 Big datan arkkitehtuuri

3.1 Yleistä

Yrityksen tai muun tahon huomatessa datan kiihtyvä ja monipuolinen kasvu on aika valmistautua big datan hallitsemiseen. Yrityksellä pitää olla tähän riittävästi laskennallista tehoa ja nopeutta sekä resurssien on tuettava kasvavia vaatimuksia. Osa datasta käsitellään heti, mutta osalle on oltava riittävästi tallennustilaa. Lisäksi mahdolliset viiveet ja häiriöajat ovat huomioitava. (Hurwitz ym. 2013, Building a Successful Big Data Management Architecture)

Sieppaa – organisoi – integroi – analysoi – toimi. Vaikka tämä kuulostaa suoraviivaiselta, ovat tiettyjen toimintojen väliset nyanssit hyvin monimutkaisia sekä tiedon validointi hyvin tärkeää. Mikäli yritys yhdistää eri datalähteitä, on tärkeää voida vahvistaa, että yhdistetyssä datassa on järkeä. Lisäksi tietoturvan ja lainsäädännön kannalta

on huomattava, että jotkin osat datasta voivat sisältää salassa pidettävää tietoa. (Mt.)

Big data -projektin aloitusvaiheessa kannattaisi miettiä seuraavia kysymyksiä:

- Kuinka paljon dataa yrityksen täytyy hallita nyt ja tulevaisuudessa?
- Kuinka usein yrityksen täytyy käsitellä dataa reaaliajassa tai lähes reaaliajassa?
- Kuinka suureen riskiin yrityksellä on varaa? Ovatko yrityksen määräykset tiukat tietoturvan ja säännösten osalta?
- Kuinka tärkeää on nopeus datan käsittelyssä?
- Kuinka varmaa ja tarkkaa datan on oltava? (Mt.)

3.2 Rajapinnat ja datalähteet

Big dataan kertyy dataa sekä sisäisesti hallitusta datasta että ulkoisista syötteistä. On oleellista ymmärtää, että big dataan kertyy paljon dataa monista eri lähteistä, ja se juuri tekeekin siitä ison. Tästä johtuen API-rajapinnat (Application Programming Interfaces) ovat big data -arkkitehtuurin ydin. (Hurwitz ym. 2013, Building a Successful Big Data Management Architecture)

3.2.1 Toiminnalliset datalähteet

Kun ajatellaan big dataa, on tärkeää ymmärtää, että siinä täytyy yhdistää kaikki datalähteet, jotka antavat täyden kuvan liiketoiminnasta. Yhdistämällä datalähteet nähdään, miten kyseinen yhdistetty data tulee vaikuttamaan liiketoiminnan suorittamiseen. Perinteisesti toimivat datalähteet ovat edelleen tarkasti strukturoitua, mutta muutosten mukana on ymmärrettävä datan sisältävän nyt laajemman datan lähteiden määrän, mukaan lukien strukturoimatonta dataa, kuten asiakastietojen ja sosiaalisen median data kaikissa sen muodoissa. (Hurwitz ym. 2013, Building a Successful Big Data Management Architecture)

Tiedonhallintaan löytyy näin ollen uusia lähestymistapoja big datan myötä. Näitä kutsutaan NoSQL- (Not only Structured Query Language) tai not only SQL-tietokannoiksi.

Yritysten täytyy siis osata kartoittaa data-arkkitehtuuri kaikille erityyppisille toiminnoille. Näin tehdessä varmistetaan, että oikea data on saatavilla oikeaan aikaan tarvittaessa. Tarvitaan myös data-arkkitehtuuria tukemaan uutta monimuotoista strukturoimatonta sisältöä varten. Big datan hallinnassa täytyy sisällyttää molemmat relaatio- ja ei relaatiotietokannat sekä strukturoimaton data saadakseen täyden kuvan yrityksen toiminnasta. (Mt.)

3.2.2 Suodatus

Tämän rajapinnan on tarkoitus suodattaa ”melu” oikeasta informaatiosta. Rajapinnan täytyy siis pystyä käsitellä suuria määriä vauhdilla tulevaa monipuolista dataa. Sillä pitäisi olla myös kyky kelpuuttaa, siivota, muuntaa, pakata sekä yhdistää dataa myöhäisempiä prosessointivaiheita varten. (Sawant ja Shah 2013, Ingestion Layer)

3.2.3 Hajautettu tallennus

Laajasti hajautettujen tallennusten ja prosessointien käyttö big datassa on yksi olennainen osa yritysten tekemistä muutoksista. Hajautettu tallennusjärjestelmä varmistaa virheensietokykyä ja rinnakkaisuus mahdollistaa nopeiden hajautettujen algoritmien prosessoinnin suuresta datan määrästä. Hadoopin hajautettu tiedostojärjestelmä HDFS (Hadoop Distributed File System) on big datan tallennuskerroksen kulmakivi. (Sawant ja Shah 2013, Distributed (Hadoop) Storage Layer)

Hadoop on avoimen lähdekoodin kehys, joka mahdollistaa valtavien data määrien hajauttamisen kustannustehokkaiden koneiden välillä. Se tarjoaa hajautettujen laskentaohjelmistojen tekniikan yhdistämisen käytössä olevaan sovelluslogiikkaan, jota halutaan suorittaa. Hadoop antaa mahdollisuuden vuorovaikuttaa loogisella prosessointiklusterilla ja tallentaa solmukohdat käyttöjärjestelmän ja keskusyksikön käytön sijasta. (Mt.)

HDFS on tiedostojärjestelmä, joka on suunniteltu tallentamaan erittäin suuria määriä informaatiota (tera- sekä petatavuja) suurelle määrälle koneita klusterissa. Se tallentaa dataa luotettavasti tallentaen koko tiedoston tai osia tiedostosta, toimii kaupallisilla laitteistoilla sekä tukee ”kerran kirjoitettua, monesti luettavaa” -mallia dataoikeuksissa. (Mt.)

HDFS vaatii kuitenkin monimutkaisen luku- ja kirjoitustiedostojärjestelmäohjelman kehittyneiltä tekijöiltä. Datan käsittely tai muuttaminen ei ole mahdollista, koska HDFS:ään ei ole mahdollista päästä käsiksi loogisen datarakenteen kautta. Helpottaakseen tätä ongelmaa, tarvitaan käyttöön uusi hajautettu NoSQL-tietokanta, jotka ovat vallitsevia big datassa. Relaatio- sekä NoSQL-tietokantojen yhdistäminen varmistaa oikean datan saamisen, kun sitä tarvitaan. (Mt.)

3.2.4 Fyysinen infrastruktuuri

Big dataa tukeva redundanttinen eli päällekkäinen fyysinen infrastruktuuri on keskeinen osa toiminnallisuutta ja skaalautuvuutta. Ilman tätä kustannustehokasta vakaata fyysisen infrastruktuurin mahdollisuutta big datasta ei todennäköisesti olisi tullut näin suurta trendiä. Tukeakseen odotuksenvastaista muotoa tai arvaamattomia määriä dataa on infrastruktuurin oltava siis erilainen perinteisiin datanhallinta ja tallennus tilanteisiin nähden. Hadoopin fyysinen infrastruktuurikerros HPIL (Hadoop Physical Infrastructure Layer) perustuu big datassa hajautettuun laskentamalliin. Hajautettu laskentamalli tarkoittaa sitä, että data voi olla fyysisesti useammassa eri paikassa. Data voidaan linkittää yhteen verkkojen välityksellä käyttäen hajautettuja tiedostojärjestelmiä ja useita eri big datan analyysityökaluja sekä sovelluksia. (Hurwitz ym. 2013, Building a Successful Big Data Management Architecture)

Redundanttisuus on tärkeää, koska käsiteltävää dataa on niin paljon eri lähteistä. Päällekkäisyys tulee ilmi monissa eri muodoissa. Mikäli yrityksellä on oma pilvipalvelu, on suotavaa, että se on rakennettu redundanttisesti, jotta se skaalautuu tukemaan vaihtelevia kuormituksia tulevaisuudessa. Jos yritys taas haluaa säilyttää oman sisäisen IT:n kasvun, voidaan käyttää ulkoisia pilvipalveluita. Ulkoisia pilvipalveluita käyttäen yritys voi säilyttää nykyisiä tai lisätä omia sisäisiä resursseja. Joissain tapauksissa tämä voi ilmetä SaaS:ina (Software as a Service) tarjoten monimutkaisia dataanalyysejä palveluna. SaaS:lla saavutetaan halvemmat kustannukset, nopeampi käyttöönotto sekä saumaton kehitys jo olemassa olevalle teknologialle. (Mt.)

3.2.5 Alustan hallinnointi

Tämä kerros tarjoaa työkalut ja tietokantojen kyselykielet HDFS:ää käyttäviin NoSQL-tietokantoihin, jotka ovat fyysisen infrastruktuurin päällä. Hallintakerros pääsee dataan käsiksi, suorittaa kieliä ja hallinnoi alempia kerroksia käyttämällä muun muassa Pig- ja Hive-ohjelmointikieliä. (Sawant ja Shah 2013, Hadoop Platform Management Layer)

3.2.6 Tietoturva ja tietosuoja

Mitä tärkeämmäksi osaksi big datan analysointi tulee yritykselle, sitä tärkeämmäksi tulee huomioida siihen liittyvä tietoturva. Esimerkiksi terveydenhoitoalalla käsitellään big data -sovelluksissa yksityissuojan alaisia tietoja, joten on erittäin tärkeää suojella potilaiden yksityisyyttä huomioiden, että kuka saa nähdä tiedot ja missä olosuhteissa he voivat niin tehdä. Yrityksen on siis pystyttävä tarkistamaan käyttäjien henkilöllisyys sekä samalla myös suojata potilaiden henkilöllisyys. Tämän tyyppiset tietoturvavaatimukset on otettava huomioon heti alussa, eikä vasta jälkikäteen välttyäkseen ongelmilta. (Hurwitz ym. 2013, Building a Successful Big Data Management Architecture)

Ilman vaadittavia oikeuksia ei-luotettavat tekijät voivat palauttaa datasta vääristyneitä tuloksia, joita ei haluta. Nämä luovat kokonaisuudesta virheellisesti muodostettuja tuloksia, jotka taas vaikuttavat haitallisesti datasta muodostettuun informaatioon. Suuresta datan määrästä tämän kaltaiset tietoturvarikkomuksen aiheuttamat poikkeamat voivat jäädä helposti huomaamatta ja aiheuttaa merkittävää vahinkoa laskennalle ja päättelylle. (Sawant ja Shah 2013, Security Layer)

NoSQL-tietokannat ovat vielä kehitykseltään alussa ja tarjoavat helpon kohteen tietomurroille. Suurten klustereiden hyödyntäminen sattumanvaraisesti ketjuissa ja big data -kokoelmien arkistoinnissa aiheuttaa helposti sen, että ei tiedetä, missä data on tallennettuna ja unohdetaan tarpeettoman datan poistaminen. Kyseisenlainen data voi näin päätyä vääriin käsiin ja aiheuttaa yritykselle vakavan tietoturvauhan. (Mt.)

Big data -projektit ovat luontaisia kohteita tietoturvaongelmille hajautetun datan, yksinkertaisten ohjelmointimallien ja avoimien palvelukehysten takia. Kuitenkin tieto-

turva täytyy toteuttaa tavalla, joka ei vaikuta heikentävästi suorituskykyyn, skaalautuvuuteen tai toiminnallisuuteen. Lisäksi tietoturvan pitäisi olla suhteellisen yksinkertainen hallita ja ylläpitää. (Mt.)

Toteuttaessa tietoturvan perustaa, lähtökohtana olisi täytettävä ainakin nämä kyseiset kohdat:

- Tietokoneiden autentikointi käyttämällä protokollia, kuten Kerberos.
- Tiedostokerroksen salaus.
- Luotettavien avainten ja sertifikaattien allekirjoitusavainten hallintapalvelu (KMS, Key Management Service).
- Työkalujen, kuten Chefin tai Puppetin käyttö datakokoelman sijoittamisen validoinnissa tai hyväksyessä korjaustiedostoja virtuaalikoneille.
- Tietokoneiden välisen kommunikointilokien tallentaminen ja hajautetun lokimekanismin käyttö poikkeavuuksien jäljittämiseen eri kerroksissa.
- Tietokoneiden välisen turvatun kommunikoinnin varmistaminen käyttäen muun muassa SSL- (Secure Sockets Layer) ja TLS-salauksia (Transport Layer Security). (Mt.)

3.2.7 Monitorointi

Monien hajautettujen datavarastointiklustereiden ja useiden datalähteiden suodatuspisteiden takia on tärkeää saada kokonaiskuva big datan teknologiapinosta, jotta päästään palvelutasosopimusten (SLA, Service-Level Agreement) määrittelemään käyttämättömyysaikaan. (Sawant ja Shah 2013, Monitoring Layer)

Monitorointijärjestelmien täytyy olla tietoisia muun muassa suurista hajautetuista klustereista, jotka ovat sijoitettu yhdistyen toisiinsa, eri käyttöjärjestelmistä ja laitteistoista, koska tietokoneiden on kommunikoitava monitorointityökaluille käyttäen korkean tason protokollia, kuten XML:ää (Extensible Markup Language) binääriformaatin sijasta. Järjestelmän pitäisi myös tarjota työkaluja datavarastoihin ja visualisointiin. Suorituskyky on monitoroinnin avainparametri, jotta kustannukset pysyvät

alhaalla ja rinnakkaisuus taas korkealla. Avoimen lähdekoodin työkalut, kuten Ganglia ja Nagios ovat laajalti käytössä big datan teknologiapinon monitoroinnissa. (Mt.)

3.2.8 Visualisointi

Suuri määrä big dataa voi johtaa informaation ylikuormitukseen. Kuitenkin, mikäli visualisointi on huomioitu ottaa mukaan jo varhaisessa vaiheessa oleelliseksi osaksi big datan teknologiapinoa, on se erittäin käytännöllinen datan analysoijille. Visualisoinnilla saavutetaan nopeammin näkemys tuloksista sekä lisätään mahdollisuuksia tarkastella eri näkökulmista dataa vaihtelevilla visuaalisilla malleilla. (Sawant ja Shah 2013, Visualization Layer)

Kehittyneitä visualisointityökaluja ovat muun muassa Tableau, Clickview, Spotfire, MapR ja Revolution R. Nämä työkalut toimivat perinteisten komponenttien, kuten raporttien, ohjauspaneelien sekä kyselyjen päällä. Tällä arkkitehtuurilla yrityksen loppukäyttäjät näkevät perinteisen liiketoiminnan datan sekä big datan yhdistettynä yksittäisenä näkymänä. (Mt.)

4 Palvelutyypit

4.1 Datapalveluiden organisointi ja työkalut

Kasvava määrä dataa tulee useista eri lähteistä. Dataa tulee muun muassa koneista, sensoreista ja valtavista julkisista sekä yksityisistä lähteistä, eikä data näin ollen ole hyvin organisoitua tai suoraviivaista. Aikaisemmin yritykset eivät ole yksinkertaisesti olleet kyvykkäitä käsittelemään dataa tai se on ollut aivan liian kallista. Vaikka datan tallennus olisikin ollut mahdollista, niin yrityksillä ei ole ollut työkaluja analysoida ja hyödyntää kyseistä dataa. Aikaisemmin vain muutamilla työkaluilla on saatu järkeä näin isoista datan määristä, mutta olemassa olevat työkalut olivat monimutkaisia käyttää, eivätkä ne tuottaneet tuloksia siedettävässä ajassa. (Hurwitz ym. 2013, Building a Successful Big Data Management Architecture)

Laskentatehon kasvaessa ja tietokoneiden komponenttien hintojen tullessa alaspäin on yrityksillä nyt mahdollisuus käsitellä tuota dataa, johon ennen vain kalliit supertietokoneet olivat kyvykkäitä. Varsinainen läpimurto big datassa tapahtui, kun yritykset, kuten Yahoo!, Google ja Facebook tulivat tilanteeseen, että syntyvälle datalle olisi pystyttävä tekemään jotakin taloudellisen hyödyn saamiseksi. Näiden yhtiöiden oli luotava uusia teknologiota saadakseen hyödyn big datasta. Heidän luomansa ratkaisut tehokkaasta ja kustannustehokkaasta datan analysoinnista ovat luoneet MapReducen, Big Tablen ja Hadoopin, jotka ovat uutta sukupolvea datanhallinnassa. (Mt.)

4.1.1 MapReduce

MapReduce on Googlen suunnittelema tapa suorittaa toimintoja sarjassa. ”Map”-komponentti jakaa ohjelmointiongelmat ja tehtävät suurilukumääräisten järjestelmien kesken. Samalla se käsittelee korvattavat tehtävät tavalla, joka tasapainottaa kuormituksen ja hallitsee palautumisen virhetilanteista. Hajautetun laskennan valmistuttua ”reduce”-toiminto kokoaa kaikki osat taas yhteen luodakseen tuloksen. (Hurwitz ym. 2013, Building a Successful Big Data Management Architecture)

MapReduce yksinkertaistaa syntyviä prosesseja, jotka analysoivat suuria määriä strukturoitua ja strukturoimatonta rinnakkaisdataa. Mahdolliset syntyvät laitteistoongelmat hoidetaan piilossa taustalla loppukäyttäjien sovellutuksille näkymättömästi. Näin pystytään tarjoamaan luotettava sekä virheitä suvaitseva valmius. (Sawant ja Shah 2013, Hadoop Platform Management Layer)

4.1.2 Big Table

Big Table on Googlen kehittelemä hajautettu tallennusjärjestelmä skaalautuvaa strukturoitua dataa varten. Data on tässä organisoitu rivillisiin ja sarakkeellisiin taulukkoihin. Perinteisiin tietokantamalleihin verrattuna Big Table on harva, hajautettu sekä vakaa moniulotteinen lajiteltu kartta (map). Se on luotu tallentamaan valtavia määriä dataa kaupallisilta palvelimilta. (Hurwitz ym. 2013, Building a Successful Big Data Management Architecture)

4.1.3 Hadoop

Hadoop on Apache-hallintainen ohjelmistokehys, joka on johdettu MapReducesta ja Big Tabletista. Hadoop mahdollistaa MapReduceen pohjautuvien sovellusten ajamisen isojen kaupallisten laitteistojen klustereissa. Hadoopin perustana on laskenta-arkkitehtuuri, joka luotiin tukemaan Yahoo!:n liiketoimintaa. Hadoop on suunniteltu rinnakkaiseen datan prosessointiin tietokoneissa. Tämä nopeuttaa laskentaa ja piilottaa mahdollisia viiveitä. Hadoopista löytyy kaksi ensisijaista komponenttia: Massiivinen skaalautuva hajautettu tiedostojärjestelmä HDFS, joka voi käsitellä petatavuja dataa sekä massiivinen skaalautuva MapReduce-kone, joka laskee tulokset sarjassa. (Hurwitz ym. 2013, Building a Successful Big Data Management Architecture)

4.2 ICT-alan suuryritysten ratkaisut

Tietotekniikka on alun perin luotu datan tallentamista sekä käsittelyä varten. Big datassa ei siis sinällään ole mistään uudesta asiasta kyse. Kyse on vain muutoksesta datan muodossa ja määrässä. Nykyään jokaisella suuremmalla ICT-alan toimijalla on omat ratkaisunsa big datan tarjoamiin mahdollisuuksiin sekä haasteisiin. Parhaimmat hyödyt yritykset saavat yleensä big data -ratkaisuista, jotka yhdistelevät tarkoitukseen sopivia ja toisiinsa sulautuvia teknologioita ja käytäntöjä. (Salo 2013, 52, 59)

4.2.1 Amazon

Amazon on tunnettu kirjojen ja nykyään myös muidenkin tuotteiden vähittäismyyntikauppiaana verkossa. Pilvipalveluiden tarjoajana se on yksi edelläkävijöistä tallenustilaa tarjoavalla S3-palvelullaan (Simple Storage Service). Muun muassa Dropbox sekä Ubuntu One käyttävät tallennuksessa S3-palvelua. Skaalaa kyseisellä palvelulla on valtavasti. Palveluun on lisätty tallennettuja objekteja yli 1000 miljardia ja niiden koon vaihdellessa yhdestä bitistä viiteen teratavuun on datan määrä valtava palvelimilla. Hinnoittelussa Amazon käyttää joustavaa hinnoittelua pelkästään tallennetun datan määrän mukaan. Hinta ensimmäiseen teratavuun asti on noin 0,07 euroa gigatavulta (GB) per kuukausi. Tämän jälkeen hinta alenee tallennettavan volyymin myötä. Tallennettua dataa turvaa kehittynyt tietoturva ja joukko sertifikaatteja, joten palvelun luotettavuus ja datan säilyvyys ovat huippuluokkaa. Amazon tarjoaa myös

Glacier-nimistä tallennuspalvelua. Tässä hinta on pudotettu 0,01 euroon gigatavulta, mutta datan saa vain pyynnöstä ladattua. Pyynnön aikaviive on noin kolmesta viiteen tuntiin ja tämän jälkeen kyseinen data on käytettävissä 24 tuntia, ennen sen uutta varastointia. (Salo 2013, 52)

Muita varteenotettavia Amazonin tarjoamia palveluita ovat Elastic MapReduce, EC2 sekä DynamoDB. Elastic MapReduce on Hadoop-klusteri, jolla voi ”louhia” dataa pilvipalveluna kahdella eri versiolla. Käytettävissä on joko avoimen lähdekoodin Hadoop-projekti tai MapR-nimisen yrityksen oma versio Hadoopista sisarprojekteineen. EC2-palvelu on AMI-virtuaalipalvelimia (Amazon Machine Image) tarjoava mahdollisuus. Valittavana on joustavasti eri kapasiteetillä varustettuja palvelimia, jokaisen asiakkaan omien tarpeiden mukaan. DynamoDB on pilvipalveluna toteutettava ei-relaatiotietokantaratkaisu. DynamoDB skaalautuu automaattisesti tarpeita vastaavaksi, kun käyttäjä itse määrittelee ensin tarvittavan suorituskyvyn. Tallennustilana käytetään nopeita SSD-levyjä (Solid-State Drive) ja kyseinen palvelu on lisäksi mahdollista integroida Elastic MapReduce -palveluun. (Salo 2013, 53)

4.2.2 EMC

EMC on maailman johtava pilvi-, big data- ja tietoinfrastruktuuritoimittaja. Vuonna 2012 yrityksen liikevaihto oli 22 miljardia dollaria. Tallennusratkaisut, ohjelmistot ja palvelut kuuluvat EMC:n tarjontaan big datan osalta. (Salo 2013, 54)

Greenplum UAP (Unified Analytics Platform) on EMC:n tarjoama laitteistoratkaisu, joka voi sisältää Greenplum tietokanta-, Hadoop- ja DIA-moduleita (Data Integration Appliance). UAP on nopeasti käyttöönotettavissa valmiiksi optimoidulla suorituskyvyllä sekä tallennuskapasiteetillä. EMC:n tarjoama tuki tulee laitteistolle ja Greenplum- sekä Hadoop-ohjelmistoille. (Mt.)

Greenplum on kolmannen sukupolven ratkaisu tietokantoihin. Se kykenee yhdistämään joustavasti strukturoitua, semi-strukturoitua ja strukturoimatonta dataa tulevia kehittyneitä analyysejä varten. Greenplumilla on mahdollista tallentaa dataa tietokantaan 13,7 teratavun tuntivauhdilla yhdellä palvelinkehikolla, joka sisältää 16 palvelinta. Lisäksi Greenplum kykenee lukemaan ja kirjoittamaan Hadoop-tietojärjestelmään suoraan ilman datan siirtoa tai muuntamista toiseen muotoon. (Mt.)

EMC tarjoaa Hadoopia kolmena eri versiona. GP HD (Greenplum Hadoop), joka on perinteinen versio ja perustuu avoimeen lähdekoodiin, yrityksille suunnattu GP HDEE (Enterprise Edition) sekä GP MR (MapR). (Mt.)

Isilon tarjoaa big datan varastoinnin jopa 20 petatavulle sekä mahdollistaa Hadoopin yhteensopivuuden korkealla käyttöasteella kustannustehokkaasti. Isilonin skaalautuvuus on joustavaa kolmesta palvelimesta aina 144 palvelimen kokoiseen klusteriin. Isilon käyttää OneFS-tiedostojärjestelmää, jonka patentoidut algoritmit mahdollistavat yli 80 % hyötysuhteen käytössä olevasta kapasiteetista vikasietoisesti. Mikäli Isilonin käytössä oleva palvelinklusteri kaipaa suorituskyvyn tai kapasiteetin osalta laajennuksia, onnistuu se ilman käyttökatkoa. Tarvittavien ominaisuuksien lisäyksen jälkeen Isilon jakaa kuormituksen automaattisesti koko järjestelmälle. (Salo 2013, 55)

DSSD D5 tehostaa tallennusta tehokkaalla IOPS:illa (Input/Output Operations Per Second), pienellä viiveellä ja suurella suorituskyvyllä. Yhdistämällä jaetut flash-muistit jokaiselle tietokoneelle kolmannen sukupolven PCIe-väylän (Peripheral Component Interconnect express) kautta pystytään maksimoimaan datasta saatavat tulokset ja suoritusteho suurille työmäärille tietokannoista, HDFS:tä tai paikallisista rinnakkaisista tiedostojärjestelmistä. (EMC - DSSD D5 2016)

4.2.3 Google

Vaikka Hadoop on alun perin Googlen luoma, niin Google ei tarjoa sitä kuitenkaan kaupallisesti. Google tarjoaa MapReduce nimellä kulkevaa ominaisuutta App Engine -pilvipalvelusta. App Engine on PaaS-ratkaisu (Platform as a Service), jonka saa maksutta käyttöön. Se tarjoaa ilmaiseksi samat resurssit, joita Google itsekin käyttää ja tarjoaa monia eri rajapintoja käytettäviksi. Näin ollen sovelluskehitys on App Enginellä nopeaa ja helppoa. Haittapuolena kuitenkin on se, että App Enginessä on heikko kontrolli käytössä olevaan infrastruktuuriin ja mahdollinen vaihto toisen palveluntarjoajan ratkaisuihin voi olla paikoitellen vaikeaa Googlen omien ratkaisujen vuoksi. Java, Python ja Googlen oma Go ovat App Enginen tarjoamat ohjelmointikielet kehitysalustoille. (Salo 2013, 56)

Amazonin S3- ja EC2-palveluiden kilpailevat versiot ovat Google Storage ja IaaS-pohjainen (Infrastructure as a Service) Google Compute Cloud pilvipalvelut. Compute

Cloud ei ole valmis Hadoop-alusta, mutta se tarjoaa mahdollisuuden erilaisiin käyttötarkoituksiin. (Salo 2013, 57)

4.2.4 IBM

IBM lähestyy big dataa pyrkien entistä älykkäämpiin analyysiratkaisuihin, joiden avulla pystytään kasvattamaan koko yrityksen analyyttistä kykyä tuottaa taloudellisesti arvokasta tietoa perinteisestä datasta sekä big datasta. IBM:n big data -analytiikka-alusta (IBM Big Data Platform) on vaatimusten mukaan mukautuva sekä skaalautuva ja sen avulla voidaan analysoida minkälaista dataa tahansa perinteisen liiketoimintatiedon lisäksi. Datan ollessa liikkuvaa tai levossa olevaa on IBM kehittänyt kaksi analytiikkaratkaisua. IBM InfoSphere Streams on tietovirtojen analysointiin kehitetty työkalu ja IBM PureData for Analytics on analyyttinen tietovarasto. (Salo 2013, 57, 61)

IBM InfoSphere Streams perustuu suurten tietovirtojen äärimmäisen nopeaan analysointiin, joissa tietoa vastaanotetaan ja välitetään eteenpäin hyödyntäen rinnakkaisia verkkokapasiteetteja. Tarvittaviin vaatimuksiin on myös tarkasti optimoituja algoritmeja, jotka pystytään hyödyntämään nopeasti ajettavalla koodilla rinnakkaisoperaatioissa. IBM InfoSphere Streams sisältää valmiita adaptereita ja algoritmeja erityyppisille datavirroille sekä myös kehitys- ja hallintaympäristön. (Salo 2013, 61)

IBM PureData for Analytics perustuu Netezzan kehittämään AMPP-teknologiaan (Asymmetric Massively Parallel Processing). Tässä big data -operaatioita pystytään suorittamaan tehokkaasti optimoidussa FPGA-ympäristössä (Field-Programmable Gate Array) lähellä tallennettua dataa. Verrattaessa ohjelmistopohjaisiin tietokantaratkaisuihin AMPP kykenee 10-100 kertaiseen suorituskykyyn sekä kykenee skaalautumaan petatavujen verran. IBM PureData for Analyticsin ratkaisut ovat kustannustehokkaita ja nopeita ottaa käyttöön. Valmiiksi integroidut palvelinlaitteistot, tietovarastot ja ohjelmistot tekevät siitä myös helposti hallittavan. (Salo 2013, 61-62)

Nämä kyseiset IBM:n ratkaisut ovat esimerkkejä äärimmäisen skaalautuvista rinnakkaisprosessointiin perustuvista ratkaisuista. Datan käsittely ja analysointi tapahtuu pienillä vasteajoilla reaaliajassa. Käytännön esimerkkejä näistä reaaliaikaisista analyy-

seistä ovat muun muassa teleoperaattorien CDR-tietojen (Call Detail Record) analysointi, kansainvälisen maksuliikenteen analysointi, säämallit ja ennusteet, liikennetietojen kerääminen, sähköverkkojen toiminta ja vikojen ennakointi sekä video-virtojen analysointi. (Salo 2013, 60-61)

Levossa olevien hajanaisten sekä monimuotoisten datavarastojen hyödyntämiseen IBM:llä on kaksi ratkaisua. IBM InfoSphere BigInsights ja IBM Social Media Analytics ovat avoimiin Hadoop- ja MapReduce-teknologioihin perustuvia ratkaisuja, joita käytetään muuan muassa terveydenhuoltoalalla, rikollisuuden ennakoinnissa ja torjunnassa, tietoturvauhkien tunnistamisessa ja vähentämisessä, uutis- ja tiedotustoiminnan tehostamisessa, kuluttajakäyttäytymisen analysoinnissa, talousarvioiden ja -ennusteiden analysoinnissa sekä energiankulutuksen pienentämisessä. (Salo 2013, 62)

IBM InfoSphere BigInsights sisältää datan keräämisen Hadoop JACL -rajapintaan (Java Command Language) toteutetuilla adaptereilla, suodattamisen ja ”louhimisen” työkalut sekä visualisoinnin. Toimintoja käytetään helpolla hieman taulukkolaskinta muistuttavalla selainpohjaisella käyttöliittymällä. Tallennettava data varastoidaan HDFS-tiedostojärjestelmään ja käsittely tapahtuu tehokkaasti MapReduce-algoritmeja käyttäen. (Salo 2013, 62-63)

4.2.5 Microsoft

Microsoft tarjoaa Hortonworksin kanssa yhteistyössä toteutetut HDInsight Hadoop -alustan, joka on tarkoitettu Windows-palvelimille, sekä Hadoop-pilvipalvelun. Microsoft Exceliin saatavilla olevat lisäosat mahdollistavat Hadoopin sekä Googlen pilvipalvelun BigQueryn käytön suurien datamäärien tallentamisessa sekä ”louhinnassa”. Excel on siis kehittymässä perinteisestä toimistotyökalusta miljardiluokan liikevaihtoa käyvien yritysten datankäsittelyn työkaluksi. (Salo 2013, 68-69)

5 Valmiit jakelut

5.1 IBM

IBM BigInsights on teollisuusstandardien mukainen Hadoop, joka tarjoaa käytettäväksi yritystason ominaisuuksilla varustettuja avoimen lähdekoodin ohjelmistoja. Se auttaa organisaatioita ja yrityksiä kustannustehokkaasti hallitsemaan sekä analysoimaan big dataa. (IBM - BigInsights for Apache Hadoop n.d.)

5.1.1 IBM BigInsights for Apache Hadoop

IBM BigInsights for Apache Hadoop tarjoaa seuraavia ominaisuuksia ja etuja:

- Kehittyneesti rakennettu analysointi Hadoop-teknologialle (IBM BigInsights Data Scientist module) vastaamaan big datan analysointi vaatimuksia ja tarpeita.
- Suunniteltu suorituskyky ja käytettävyys (IBM BigInsights Analyst module). Optimoitu suorituskykyinen kapasiteetti, visualisointi, laajat kehittäjätyökalut ja tehokkaat analytiikka toiminnot.
- Hallinta, tietoturva ja luotettavuus (IBM Enterprise Management module). Tukee ja nopeuttaa suuri skaalaisia käyttöönottoja.
- Integroituu IBM:n ja muiden informaatioratkaisuiden kanssa helpottaen datan käsittelyä ja hallintatehtäviä. (IBM - BigInsights for Apache Hadoop n.d.)

IBM:n avoimelle sovellusalustalle toteutettu BigInsights päivitetään säännöllisesti, joten se pystyy tarjoamaan uusimmat ja parhaimmat versiot Apache Hadoopin komponenteista mukaan lukien muun muassa Ambari, YARN, Spark, Knox, HBase, Hive sekä kryptattu eli salattu HDFS. Lisäksi BigInsights mahdollistaa korkea arvoisten Hadoop-analytiikkatyökalujen, kuten Big SQL, BigSheets, Text Analytics, Big R ja koneoppimisen käytön nopeuttamaan datasta saatavan tiedon ymmärtämistä. (Bluemix - BigInsights for Apache Hadoop 2015)

Avainkomponentteja, mukaan lukien infrastruktuuri, monitoroidaan ennakoivasti kellon ympäri IBM:n toimesta. Kriittiset turvallisuus korjaukset, päivitykset, muutostiedostot ja virheratkaisut sovelletaan viipymättä klustereihin. (Mt.)

IBM:n pilvitoimintoryhmä vastaa seuraavista palveluista:

- Tarjoaa ja hallinnoi palvelimet, tallennustilan sekä verkkoinfrastruktuurin klustereille.
- Tarjoaa alustavan konfiguroinnin IBM:n avoimen sovellusalustan komponenteille sekä kaikille käyttöön valituille BigInsights-moduuleille.
- Tarjoaa ja hallinnoi internetrajapinnan sekä sisäisen palomuurin suojauksen ja eristämisen.
- Monitoroi ja hallinnoi seuraavia palveluiden komponentteja:
    - Verkkokomponentit
    - Palvelimet sekä niiden paikallinen tallennustila
    - Käyttöjärjestelmät
    - Hadoop-avainten hallintapalvelimet (KMS)
    - Hadoop-klustereiden hakemistopalveluiden verkkoprotokolla LDAP (Lightweight Directory Access Protocol)
    - Ambari-klusterihallinta
- Tarjoaa ylläpidon korjaustiedostoille, mukaan lukien käyttöjärjestelmälle sopivat tietoturvan korjaustiedostot, IBM:n avoimelle sovellusalustalle sekä kaikille valituille BigInsights-moduuleille. Ylläpitoa ei kuitenkaan tarjota millekään ohjelmistolle tai komponentille, jotka käyttäjä on itse erikseen lisännyt. (Mt.)

Ylläpidon ulkopuolelle eli käyttäjän vastuulla olevat asiat:

- IBM:n avoimen sovellusalustan komponenttien, jotka ajetaan Ambarin alla, monitorointi, konfigurointi ja hallinta. Käyttäjä voi joustavasti valita mitä

komponentteja ajaa, mutta on vastuussa näiden käynnistämisestä, monitoroinnista ja pysäyttämisestä.

- Käyttäjien ja ryhmien lisääminen klustereihin.
- Palvelun ohjelmien ja sovellusten kehittäminen tarvittaessa, jotta dataa voidaan analysoida ja siitä saadaan ymmärrettävää tietoa. Näiden ohjelmien ja sovellusten laadun sekä tehokkuuden varmistaminen on käyttäjien vastuulla.
- Ylläpitämään IBM:n sallimia ohjelmistoja tai dataa, jotka käyttäjä on itse lisännyt klusteriin. IBM voi antaa tukea, mutta ei ylläpidä, siirrä tai poista mitään ohjelmistoa tai dataa, joka vaikuttaa palvelun toimivuuteen.
- Hadoop-datan kryptaus ja sen käyttö.
- Datan varmuuskopiointi ja palauttaminen, metadata, konfigurointitiedostot ja alustan parametrit.
- Yhteensopivuuden ja suorituskyvyn varmistaminen, kun sallittuja ohjelmistoja asennetaan tai IBM:n avoimen sovellusalustan komponentteja sekä BigInsights-ohjelmistoja päivitetään. (Mt.)

5.1.2 IBM BigInsightsin versiot

IBM BigInsights Quick Start Edition on IBM:n avoimelle alustalle tehty yritystason ominaisuuksilla varustettu ohjelmisto. Se sisältää visualisoinnin, tarkkailun sekä IBM BigInsights Data Scientist ja IBM BigInsights Analyst -ratkaisujen kehittyneet analysointi mahdollisuudet big datalle. (IBM - BigInsights Quick Start n.d.)

Docker Image:

Vaaditut laitteistovaatimukset yhden tai usean tietokoneen Docker Image -versiolle:

- Minimissään 12 gigatavua keskusmuistia (RAM, Random Access Memory)
- Neliydinprosessori
- 50 gigatavua vapaata lukumuistia (ROM, Read Only Memory)
- Käyttöjärjestelmä Red Hat Enterprise Linux (RHEL) 7.x - 64-bit (Docker 1.8.1)

Natiiviasennus:

Vaaditut laitteistovaatimukset ohjelmiston natiiviasennukselle, joka sisältää IBM Open Platform with Apache Hadoopin sekä Quick Start Edition for the IBM BigInsights Data Scientist Module asennukset:

- Minimissään 24 gigatavua RAM-muistia
- Minimissään 80 gigatavua ROM-muistia
- Käyttöjärjestelmä x86 tai Power 64-bit Red Hat Linux

VM Image:

Vaaditut laitteistovaatimukset ohjelmiston VM Image (Virtual Machine) -versiolle:

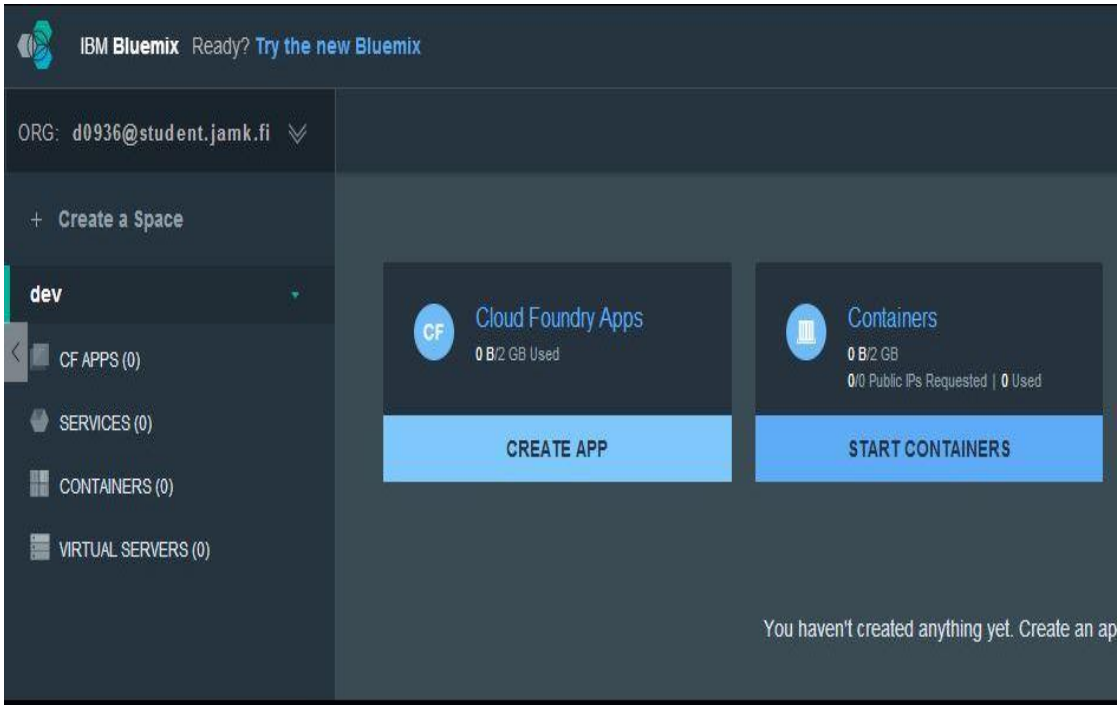
- Minimissään 12 gigatavua RAM-muistia
- Neliydinprosessori
- 50 gigatavua vapaata ROM-muistia
- Käyttöjärjestelmä VMware Windows tai VMware OS X (Mt.)

5.1.3 IBM BigInsights on Cloud

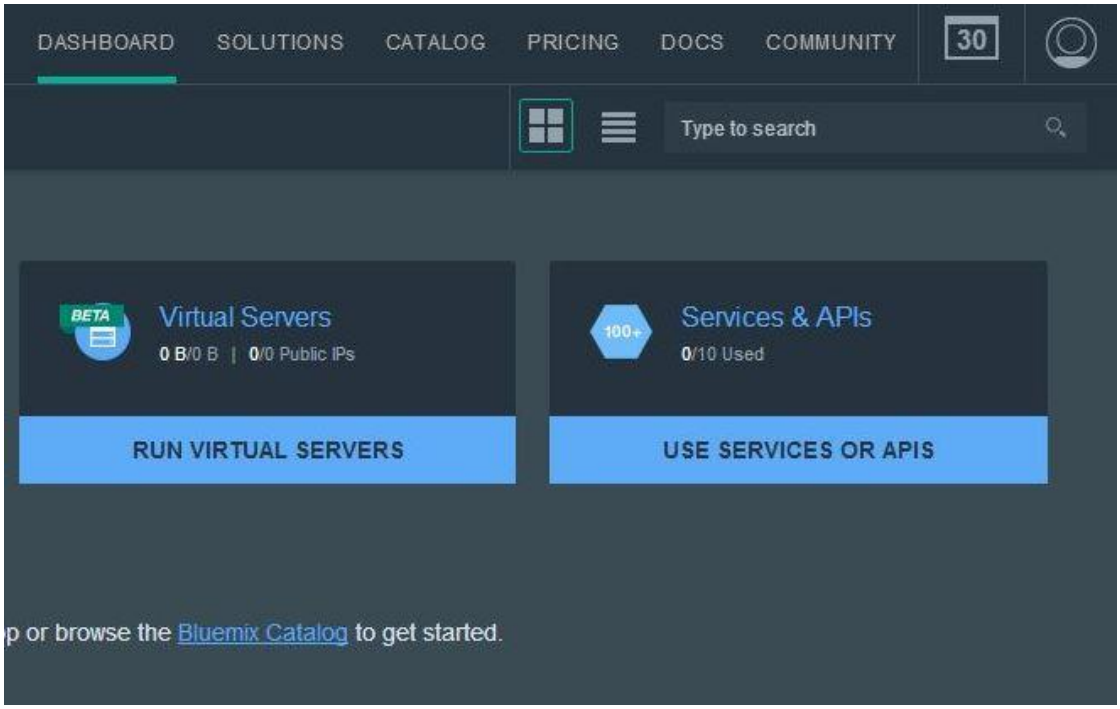
IBM Bluemixin isännöimä palvelu IBM BigInsights on Cloud on nopea ja ilmainen tapa tutustua IBM BigInsights Quick Start Editionin mahdollisuuksia pilvipalvelussa. Pilvipalvelu mahdollistaa kokeilun ilman asentamista, konfigurointia tai ylläpitoa. IBM BigInsights on Cloud -palvelun avulla luodaan pääsy Hadoop-klustereihin, rakennetaan sovelluksia ja analysoidaan strukturoitua sekä strukturoimatonta dataa. Tulosten visualisointi taulukoihin ja graafisiin kuvaajiin onnistuu myös palvelun kautta. Oman datan pystyy tuomaan Hadoopiin analysoitavaksi käyttämällä Big SQL, BigSheets, Text Analytics, Big R tai koneoppimisen ratkaisuja. (Bluemix - Analytics for Hadoop 2015)

Palvelun käyttöönotto vaatii rekisteröitymisen IBM Bluemixiin. Nimen, puhelinnumeron, maan ja sähköpostin ilmoittamisen jälkeen ilmoitettuun sähköpostiosoitteeseen tulee vielä varmistusviesti, jolla käyttäjätili varmennetaan. Ilmaisella kolmenkymme-

nen päivän kokeiluversiolla saa käyttöönsä IBM Bluemix -alustan sovellusten rakentamiseen, kaksi gigatavua suoritusmuistia (runtime) ja kaksi gigatavua tallennustilaa sekä oikeuden kymmeneen IBM Bluemix -palveluun ja API:iin (ks. kuvio 1 ja 2). Ilmaisversioon kuuluu lisäksi 50 gigatavun HDFS-tallennustila, mutta ei datan varmuuskopiointia eikä palvelutasosopimusta. Lisäksi IBM Bluemix -sovellusten yhteydet HDFS-dataan on rajoitettu kahteen sovellukseen kerralla. Mikäli IBM BigInsights on Cloudiin aikoo rakentaa ja liittää omia sovelluksia, tarvitaan myös ympäristöön vaadittavat ajurit omalle tietokoneelle asennettuna. Pilvipalvelu ei siis ole täysin puhtaasti palvelimilla toimiva kehitystyötä tehdessä. (Mt.)



Kuvio 1. IBM Bluemixin hallintapaneeli osa 1 (Bluemix - Dashboard 2016)



Kuvio 2. IBM Bluemixin hallintapaneeli osa 2 (Bluemix - Dashboard 2016)

5.2 Cloudera Enterprise

5.2.1 Cloudera Distribution Including Apache Hadoop - CDH

Cloudera Distribution Including Apache Hadoop (CDH) on Clouderan avoimen lähdekoodin Apache Hadoop -jakelu. Se sisältää kaikki johtavat Hadoop-komponentit tallentamiseen, prosessointiin, tiedon löytämiseen, mallintamiseen sekä rajoittamattoman datan käyttöön. CDH on suunniteltu täyttämään korkeimmat yrityspuolen standardit vakaudessa ja luotettavuudessa. (Cloudera - Apache Hadoop n.d.)

CDH perustuu täysin pitkäaikaisiin avoimiin standardeihin. Avoimien standardien etuna Hadoopissa on, että Cloudera pystyy tarjoamaan uudet avoimen lähdekoodin ratkaisut omalle alustalleen. Näistä esimerkkeinä mainittakoon Apache Spark, Apache HBase ja Apache Parqueet, jotka ovat ajan myötä otettu Hadoopiin mukaan pysyvästi. (Mt.)

5.2.2 CDH:n komponentit

Clouderan avoimen lähdekoodin alustan jakelu sisältää 13 avainkomponenttia käyttövalmiina. Cloudera on luonut toimivan ja kehittyneen järjestelmän, joka auttaa selviytymään big datan haasteista. Avainkomponentit ovat Clouderan tukemia CDH:n integroituja osia. (Cloudera - Key CDH Components n.d.)

- Apache Hadoop (Ydin)
  Apache Hadoopin ydinkomponentit ovat HDFS, MapReduce ja YARN. Ydinkomponentit mahdollistavat monimuotisen datan rajoittamattoman määrän tallentamisen ja prosessoinnin yhdellä sovellusalustalla.

- Apache HBase
  HBase on skaalautuva tietue- ja taulukkotallennusväline reaaliaikaisella luku- sekä kirjoitusoikeudella.

- Impala
  Impala on Hadoopin natiivi analyyttinen tietokanta. Impala mahdollistaa pieniviiveiset tietokantakyselyt useiden käyttäjien kuormituksesta huolimatta.

- Apache Sentry
  Hienojakoinen rooleihin pohjautuva auktorisointityökalu Impalalle ja Hivelle.

- Apache Sqoop
  Datan siirtokone Hadoopin integroimiseen relaatiotietokantoihin.

- Apache Accumulo
  Tietoturvallinen hajautettu tiedontallennusväline korkeaa suorituskykyä vaativille big data -sovelluksille.

- Apache Hive
  SQL-tyylinen kehys, joka sisältää metadata tietolähteen Hadoop-datan eräprosessoinnille (batch processing).

- Apache Kafka
  Hadoopille suunniteltu joustava ja tietoturvallinen julkaisu-tilaus viestintäjärjestelmä.

- Cloudera Search
  Apache Solr -käyttöinen tekstihaku, joka mahdollistaa helposti käyttäjien pääsyn Hadoopin dataan.

- Apache Flume
  Työkalu loki- ja tapahtumadatan sekä reaaliaikaisen suoratoiston keräämiseen ja yhteen kokoamiseen Hadoopiin.

- HUE
  Laajennettava graafinen verkkokäyttöliittymä, joka helpottaa Hadoopin käyttäjien tuottavuutta.

- Apache Pig
  Korkeatasoinen datavirtakieli Hadoop-datan prosessointiin.

- Apache Spark
  Avoimen standardin reaaliaikainen eräprosessointityökalu kehittyneeseen analysointiin. (Mt.)

5.2.3 CDH:n versiot

Cloudera QuickStart sisältää kokonaisen Hadoop-klusterin Docker Image tai Virtual Machine -muodossa. Mukana tulee myös Cloudera Manager -ohjelma hallintaa varten. Cloudera mahdollistaa näin ilmaisen tavan tutustua jakeluun ideaalisella ympäristöllä oppimiseen, uusien ideoiden testaamiseen ja omien sovellusten demoamiseen. Kyseiset lataukset ovat vain henkilökohtaista ja demokäyttöä varten, eikä niitä voida käyttää yrityksen käyttöönottopisteenä tuotantoklustereissa. (Cloudera - Director n.d.)

Virtual Machine Image:

- 64-bittinen isäntäkoneen käyttöjärjestelmä ja virtualisointituote, joka tukee 64-bittistä vieraskäyttöjärjestelmää

- VMwaren käyttö vaatii

  - WorkStation 8.x tai uudempi

  - Player 4.x tai uudempi

        - Fusion 4.x tai uudempi
        - WorkStationin vanhempia versioita voidaan käyttää luomalla uusi VM samalla virtuaalilevykkeellä (VMDK, Virtual Machine Disk), mutta osa toiminnoista VMware Toolsista eivät ole käytettävissä.
    - Tarvittava RAM-muistin määrä vaihtelee valitun laskentatehon mukaan
        - CDH 5 (oletus) 4+ gigatavua
        - Cloudera Express (ilmainen) 8+ gigatavua ja vähintään kaksi virtuaaliprosessoria
        - Cloudera Enterprise (60 päivän kokeiluversio) 10+ gigatavua ja vähintään kaksi virtuaaliprosessoria

Cloudera QuickStart VM on saatavilla VMware-, KVM- ja VirtualBox-formaateille. (Mt.)

CDH 5.6.0:

- Vähintään 64 gigatavua RAM-muistia. Tarvittavan muistin määrä määräytyy tarvittavan laskentatehon perusteella.
- Vähintään 500 gigatavua ROM-muistia
- Neliydinprosessori
- CDH tukee vain 64-bittisiä käyttöjärjestelmiä tietyin versiorajoituksin
    - RHEL compatible, CentOS, Oracle Linux, SUSE Linux, Ubuntu, Debian
- Vähintään kaksi virtuaaliprosessoria (Mt.)

5.2.4 Cloudera Director

Cloudera Director on tuotantovalmis Apache Hadoop -pilvipalvelu, jossa on joustava itsepalveluna toimiva käyttöönotto. Director on suunniteltu laajennettavan ohjelmistonkehyksen myötä, johon yhteistyökumppanit voivat saumattomasti integroitua. Tällä hetkellä Director sisältää integroinnit Amazon Web Servicesin (AWS) ja Google

Cloud Platformin (GCP) kanssa. Intuitiivisen käyttöliittymän kautta useat käyttäjäryhmät voivat nopeasti ottaa hyödyn pilvipalvelun kasvattamasta nopeudesta ja joustavuudesta. Käyttöönotot ovat valmiiksi konfiguroitu, joka mahdollistaa tuotteen ottamisen heti käyttöön huolimatta pilviympäristöstä. (Cloudera - Director n.d.)

Directorissa on yksinkertainen pilvikeskeinen hallinta, joka tarjoaa yksittäisen ruutunäkymän kaikista käytössä olevista pilvipalveluista. Director on vahvasti integroitu Cloudera Managerin kanssa suoria yhteyksiä varten sekä yhtenäiseen ylläpitoon klusteritason hallinnoinnissa ja monitoroinnissa. Directorin käyttöliittymä tarjoaa yksinkertaistetun hallinnan koko klusterin elinkaaren ajaksi. Ohjatuilla toiminnoilla kiihdytetään, skaalautetaan, päätetään ja jopa kloonataan klusterit tarvittaessa. Directorin ongelmatilanteita tutkii ja tukee Clouderan ammattilaisten ryhmä kellonympäri. (Mt.)

Cloudera Directorin saa käyttöönsä perustietojen antamisella rekisteröinnin yhteydessä. Directoria voi suorittaa verkkosovelluksena tai asiakasversiona tietokoneelta. Valittavissa on AWS Quick Start, asiakas- ja palvelinversiot. AWS:n sekä Google Compute Enginen palvelinkoneille on omat versionsa. (Cloudera - Director 2.0 n.d.)

Cloudera Director 2.0.0:

- Vähintään 4 gigatavua RAM-muistia.
- Vähintään 8 gigatavua ROM-muistia
- Kaksiydinprosessori
- Director tukee vain 64-bittisiä käyttöjärjestelmiä tietyin versiorajoituksin
    - RHEL ja CentOS 6.5, 6.7 ja 7.1, Ubuntu 14.04
- Cloudera Manager ja CDH: 64 GB RAM, 500 GB ROM, neliydinprosessori (Mt.)

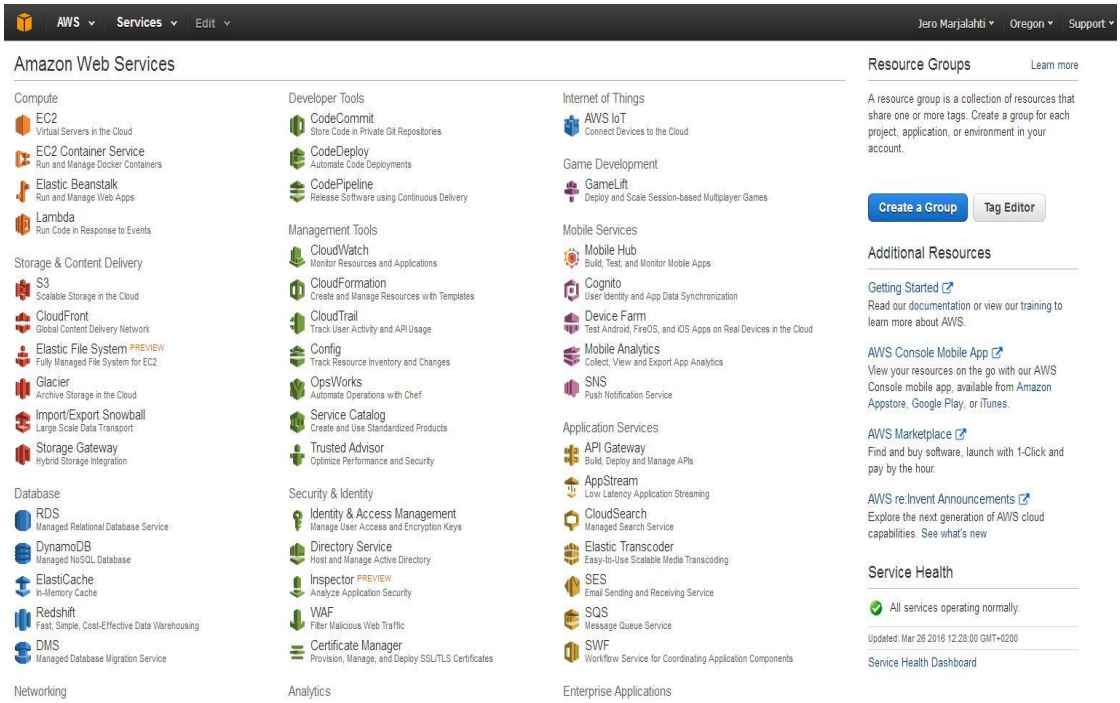
Cloudera Directorin ja AWS:n integraatiolla (ks. kuvio 3 ja 4) voidaan toteuttaa 12:n tietokoneen kokoinen klusteri. AWS Quick Startin saa käyttöönsä yhdeksi vuodeksi ilman veloitusta. Vuoden mittainen kokeilujakso sisältää tallennustilaa Amazon S3 -palvelussa 5 gigatavua, laskentatehoa Amazon EC2 -palvelussa 750 tuntia kuukaudessa, Amazon relaatiotietokantapalvelua (RDS, Relational Database Service) 750 tuntia kuukaudessa sekä 25 gigatavua tallennustilaa Amazon DynamoDB:ssä sisältäen

200 miljoonan kyselyn tekemisen kuukaudessa. Palvelun käyttöönotto vaatii perustietojen luovuttamisen lisäksi voimassaolevan luottokortin tietojen antamisen, vaikka peruspalvelu onkin ilmainen. Käyttöönotto varmennetaan sähköpostin lisäksi myös puhelinsoitolla, johon syötetään Amazonilta saatu pin-koodi. (AWS Amazon - Quick Start n.d.)

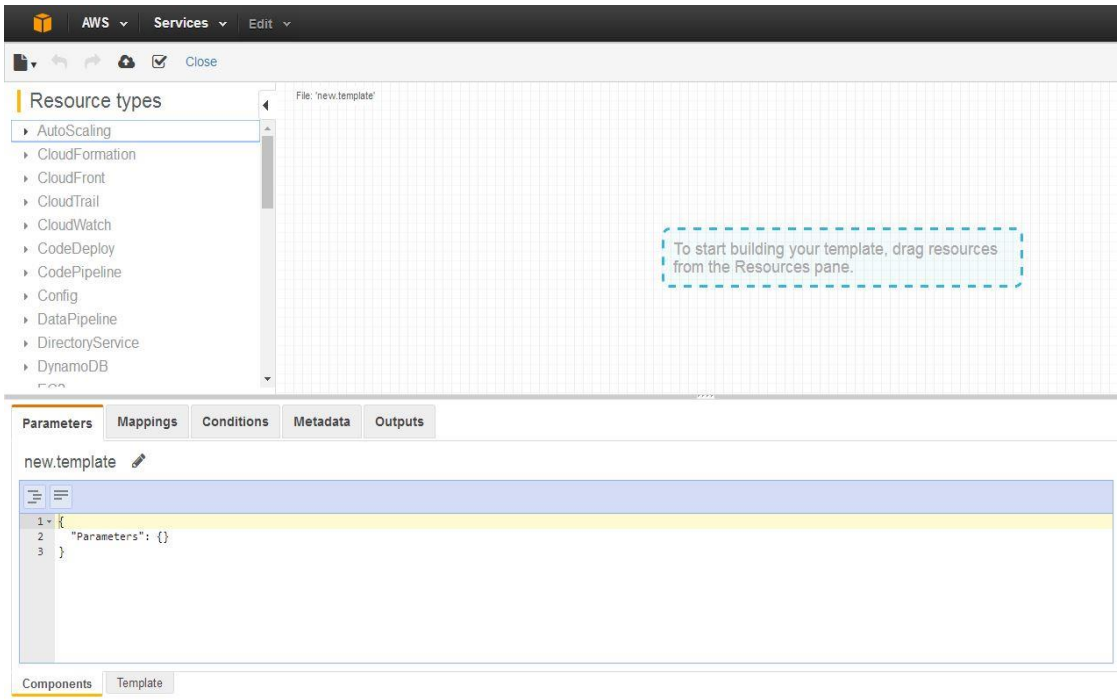
Valittavissa olevat palvelutasot:

- Perus (ilmainen)
  Asiakaspalvelu vastaa vain tili- ja laskutusasioista sekä lähteistä, jotka eivät läpäise järjestelmän tietoturvatarkastuksia. Pääsy AWS-yhteisön foorumeille.
- Kehittäjä (44€/kk)
  Teknisten kysymysten kysely ja vastaus verkkokyselyihin seuraavan 12 tunnin aikana paikallista virka-aikaa.
- Yritys (alkaen 89€/kk)
  Reaaliaikainen tuki vuorokauden ympäri puhelimitse ja chatissa, vastaus tunnissa verkkokyselyihin sekä apu kolmannen osapuolen sovellusten kanssa. Pääsy AWS:n luotettuihin neuvoihin, jotka auttavat lisäämään tehokkuutta, virheensietoa, tietoturvaa ja mahdollisesti myös rahan säästöä.
- Yhtiö
  Vastaus 15 minuutissa verkkokyselyihin. Määrätty tekninen yhteyspäällikkö, joka on asiantuntija yhtiön asioissa. Hienovarainen asioiden hoito kriittisten ongelmien suhteen, joista menee ilmoitus aina tekniselle yhteyspäällikölle ja tekniselle palveluryhmälle. Hinta sovitaan erikseen yhtiön tarpeiden määrittelyn mukaan. (AWS Amazon - Sign Up n.d.)

Käyttöönottoa on helpottamassa 13 kymmenen minuutin mittaista esittelyä kuvallisilla ohjeistuksilla. (AWS Amazon - Getting Started n.d.)



Kuvio 3. Amazon Web Services -konsolin kotinäkymä (AWS Amazon - Console Home 2016)



Kuvio 4. Amazon Web Services Cloud Formation (AWS Amazon - Cloud Formation 2016)

5.2.5 Enterprise Data Hub - EDH

Palvelusopimuksen mukaan toimivan teknisen tuen lisänä Cloudera tarjoaa etuna ennakoivan ja proaktiivisen tuen mahdollisuudet. Enterprise Data Hub (EDH) perustuu asiakkaiden tuntemiseen. Clouderan proaktiivinen tukiyksikkö varmistaa, että asiakkaat hyötyvät jokaisesta sopimuksen elementistä heti käyttöönotosta lähtien. Tuotteiden mukana kulkeva prosessi tarkkailee käyttäjien tarvitsemaa teknistä avustamista, esittelee avaintuotteiden dokumentaatiot ja yhteisön resurssit sekä varmistaa, että käyttäjät pystyvät ottamaan täyden hyödyn verkossa toimivasta tukiportaalista saavuttaakseen yritystoimintansa tavoitteet. Proaktiivinen tuki sisältää myös tunnettujen konfiguraatio-ongelmien läpikäynnin ja tarjoaa käyttökuvioiden vertailun tehostaakseen käyttäjien toimintoja sekä tulevaisuuden muutosten suunnittelua. (Cloudera - Predictive and Proactive Support n.d.)

Clouderan käyttäjät hyötyvät sopimukseen kuuluvasta analysoinnista, jossa hyödynnetään kymmenien tuhansien tietokoneiden suorituskykyä. Clouderan big datasta kerättävä ennakoiva tukimalli kerää yhteismuotoista dataa käyttöönotetuista Hadooptapahtumista, kaappaa tietoa meneillään olevista tukitehtävistä ja parittaa niitä Apache-yhteisön luomien viimeisintä tekniikkaa edustavien ratkaisujen kanssa. Näin asiakkaita autetaan minimoimaan ongelmat jo ennen niiden ilmenemistä. Cloudera on sisäisellä HBase-klusterillaan saavuttanut tukipyyntöjen ratkaisujen käsittelyajan pienentämisen 35 %:lla. (Mt.)

Proaktiivisen tuen avainkomponentit:

- Mukana kulkeva tukiprosessi ja tukiprosessiin pääsy.
- Lisenssiavainten varaukset.
- Sovellusalustan käytön raportointi.
- Yhteismuotoisten suorituskykyjen analysointi.
- Teknisen tuen varautuminen ennakkoon asiakkaiden päivitystarpeissa.
- Tunnettujen konfiguraatio-ongelmien läpikäynti. (Mt.)

Proaktiivisen tuen avainhyödyt:

- Valmius tehdä työtä heti haluttujen menestyskriteerien mukaisesti.
- Tukiresurssityökalujen ymmärtäminen ja hyödyntäminen.
- Sovellusalustan käytön raportointi.
- Käyttöasteen optimointi vertailussa käytettyjen analyysien perusteella.
- Reagointi aikaisessa vaiheessa mahdollisiin ongelmiin.
- Yleisten konfiguraatio virheiden välttäminen. (Mt.)

5.3 Pivotal

Pivotal Big Data Suite tarjoaa laajan ja modernin data-arkkitehtuurin, jota voidaan käyttää myös julkisissa pilvipalveluissa. Se sisältää tarvittavat osat eräprosessoinnin ja suoratoiston analysointiarkkitehtuuriin. Tuote on yhteensopiva kaikkien ODPi (Open Data Platform intiative) Hadoop -jakeluiden kanssa. Kaikki komponentit ovat avoimen lähdekoodin jakeluprojekteja tai ne ovat prosessissa tulossa sellaisiksi. Big Data Suite -tuotteiden sopimukset ovat 1-3 vuoden mittaisia ja valittavissa on rajaton kapasiteetti oman tarpeen mukaan. (Pivotal - Big Data Suite n.d.)

Pivotal Big Data Suite tarjoaa käyttöön joustavasti valittavia varmennettuja avoimen lähdekoodin ratkaisuja ja skaalautuvia tietovarastoja. Käytettävissä on muun muassa Pivotal Greenplum, Pivotal HDB ja Pivotal GemFire. (Mt.)

Pivotal Greenplum on massiivinen avoimen lähdekoodin rinnakkaistietovarasto. Greenplum on kehittynyt ja täysin ominaisuuksin varustettu. Se tarjoaa tehokkaan ja nopean analysoinnin petatavujen kokoisista datamääristä. (Pivotal - Greenplum n.d.)

Pivotal HDB on Apache HAWQ:iin perustuva Hadoop natiivi SQL-kone. HDB:n rinnakkaisprosessiarkkitehtuuri tuottaa korkean suorituskyvyn ja lähes reaaliaikaisen pieniviiveisen kyselyvasteen. Joustava SQL-kyselykone yhdistää MPP-pohjaisen (Massively Parallel Processing) analysointi suorituskyvyn, vakaan ANSI (American National Standards Institute) SQL-92, -99, ja -2003 määräystenmukaisen SQL:än sekä Apache MADlibin. Tämä mahdollistaa nopeiden ad hoc -kyselyjen ajamisen sekä nopeiden ennustavien analysointien suorittamisen. (Pivotal - HDB n.d.)

Pivotal GemFire on Apache Geodeen pohjautuva skaalautuva dataverkko. GemFiren avulla voidaan luoda sovelluksia, jotka toimivat reaaliaikaisesti hajautetun teknologian ansiosta. Sovellukset saadaan skaalautumaan joustavasti oletusten mukaisesti tai yllättävien kapasiteettia vaativien piikkien aikana. (Pivotal - Gemfire n.d.)

5.3.1 Pivotal Big Data Suiten komponentit

Pivotal on lisännyt kaksi komponenttia tuomaan lisää arvoa. Spring XD ja Apache MADlib -komponentit helpottavat modernin data-arkkitehtuurin käyttöönottoa ja hallitsemista. Lisäksi komponentit tukevat nopeaa ja joustavaa datan käsittelyä sekä koneoppimisen kirjastoja skaalautuville järjestelmille. (Spring - Project Spring XD)

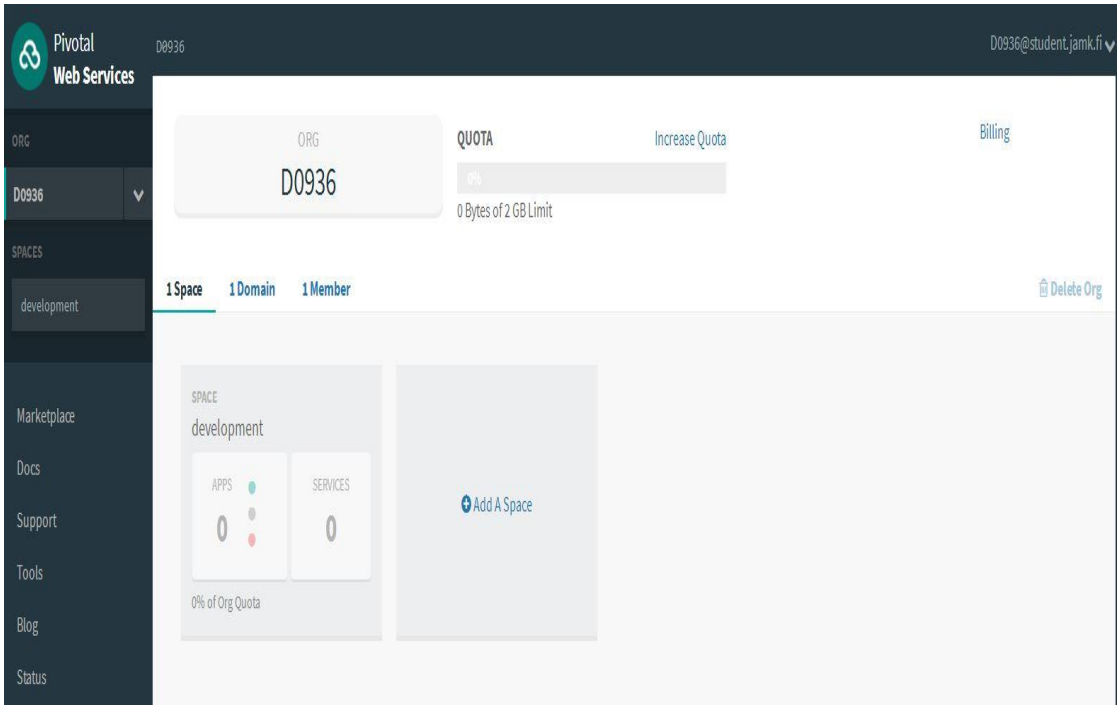
Spring XD on yhtenäinen skaalautuva avoimen lähdekoodin kehysjakelu. Se on hajautettu ja laajennettava järjestelmä datan suodatukselle, reaaliaikaiselle analyysille, eräprosessoinnille sekä datan viennille. Spring XD -projektin tarkoitus on yksinkertaistaa big data -sovellusten kehitystä, mutta kuitenkin olla rajoittamatta yrityksen valintoja tarpeiden mukaan. Uudet ohjelmistokehittäjät voivat käyttää konfigurointikäyttöistä työkalua Spring XD -sovellusten kehittämisessä ohjelmoinnin sijasta. Java-ohjelmistokehittäjät voivat myös helposti laajentaa sovellusalustaa tai DSL:ää (Domain-Specific Languages) tutuilla testi- ja automaatiotyökaluilla, jotka ovat periytyneet Spring Batchista ja Spring Integrationista. (Mt.)

Apache MADlib on tehokas avoimen lähdekoodin kirjasto skaalautuville koneoppimisen tietokanta-algoritmeille. Koneoppiminen ei ole uusi asia, mutta räjähtävällä nopeudella kasvavan datan määrien ja lähteiden myötä siitä on tullut kriittinen komponentti big data -analytiikassa. Koneoppimisen algoritmit eivät pelkästään mahdollista toistuvien kuvioiden ja kehityssuuntien tunnistamista big datasta, vaan myös mahdollistaa korkea-arvoisten ennusteiden käytön päätösten ja toimien suhteen lähes reaaliaikaisesti ilman ihmisten puuttumista prosessiin. Koneoppimisen analytiikkapaketit ovat kehittyneet ulkoisiksi sovellusalustoiksi, joita usein suoritetaan suurten tietolähteiden, kuten MPP-tietovarastojen tai tuotannon Hadoop-järjestelmien ulkopuolelta. Apache MADlib tarjoaa rinnakkaisdata toteutuksia koneoppimiselle, matemaattisia ja tilastollisia menetelmiä Pivotal Greenplumpille, PostgreSQL:lle ja Apache HAWQ:lle. MADlib käyttää MPP-arkkitehtuurin täyttä laskentatehoa hyödyksi erittäin

suurten datamäärien analysoinnissa, kun kilpailijoiden toteutuksissa rajoitettu datan määrä ladataan yksittäisen tietokoneen muistiin. (Pivotal - MADlib n.d.)

5.3.2 Pivotal Cloud Foundry - PCF

Pivotal Cloud Foundryn (PCF) käyttöönotto vaatii perustiedoilla rekisteröitymisen ilmaiseen 60 päivän kokeilujaksoon. Rekisteröitymisen varmennus tapahtuu sähköpostilla sekä tekstiviestillä. Kokeilujakso sisältää kaksi gigatavua tallennustilaa. Käyttöönottoa on tehty helpottamaan noin 15 minuutin mittainen opastus vaihe vaiheelta. Opastus sisältää esittelyn, Cloud Foundry -komentokehotekäyttöliittymän (CF CLI, Cloud Foundry Command-Line Interface) asennuksen, mallisovelluksen käyttöönoton, lokitiedostojen tarkastamisen, yhdistämisen tietokantaan sekä sovellusten skaalauksen (ks. kuvio 5). (Pivotal - Getting Started With Pivotal Cloud Foundry n.d.)



Kuvio 5. Pivotal Web Servicesin ohjauspaneeli (Pivotal - Web Services Dashboard 2016)

5.4 Hortonworks

Hortonworks luottaa 100 %:sesti avoimen lähdekoodin ratkaisuihin. Avoimen yrityspuolen Apache Hadoopin (Open Enterprise Apache Hadoop) kasvava ratkaisuiden kategoria skaalautuu asiakkaiden big data vaatimusten ja tarpeiden mukaan. Avoimen yrityspuolen Apache Hadoop korvaa aikaisempien sukupolvien Hadoop-ratkaisuiden puutteet, jotka haittasivat käyttäjiä. Varhaiset yritykset nojautuivat aikaisempiin Hadoop-projektien laajennusosiin sekä seurasivat haarautuvia lähestymistapoja. Tämä sulki pois avoimen yhteisön luomat myöhemmät innovaatiot. Lähestymistapa usein myös sitoi käyttäjät ohjelmistotoimittajakohtaisiin analyyseihin ja heikensivät myöhempiä integraatioita. Omien laajennusosien luomisen sijaan big data -kategorian ohjelmistotoimittajat luottavat yksinomaan avoimen lähdekoodin komponentteihin ja avoimeen yhteisöön. Tuloksena tästä kansainvälisestä strategiasta on avoimen yrityspuolen Apache Hadoop -ratkaisut: (Hortonworks - Hortonworks Data Platform n.d.)

- Avoimen lähdekoodin kehityksen vipuvoima.
  Tämän ansiosta yritykset hyötyvät viimeisimmistä yhteisön innovaatioista, heti niiden kypsyttyä yritys-standardien tasolle.

- Vakaat datatallennukset keskitetyillä YARN-pohjaisilla datavarastoilla.
  Käytettävissä on useita epäyhtenäisiä tapoja datan käsiksi pääsyyn tukemaan useita käyttäjiä kerralla sekä skaalaamaan tehtäviä hallitsemaan petatavuja dataa. Avoin yrityspuolen Apache Hadoop myös varmistaa täyden yhteensopivuuden Hadoopin ytimen yli edistämällä avointen standardien käyttöä.

- Kestävien operaatioiden tarjoaminen, tietoturva ja hallinta mahdollisuudet.
  Valmis yritys-standardien mukainen sovellusalusta, jonka toimivuus taataan täsmällisten ja jatkuvien projektien myötä. (Mt.)

5.4.1 Hortonworks Data Platform - HDP

Hortonworks Data Platformin (HDP) arkkitehtuuri on kehitetty ja rakennettu täysin avoimeksi. HDP tarjoaa yrityksille valmiin uuden sukupolven datasovellusalustan käyttöönoton. YARN:in ympärille rakennettu arkkitehtuuri tarjoaa datasovellusalus-

tan moninkertaisille dataprosessoinnin työmäärille hyödyntäen kaikkia prosessointimenetelmiä eräprosessoinnista, interaktiiviseen sekä reaaliaikaiseen prosessointiin. HDP tukee tärkeimpiä yrityspuolen datasovellusalustan vaatimuksia mukaan lukien hallinta, tietoturva ja operaatiot. Nämä vaatimukset kattavat eri komponettiteknologien väliset toiminnallisuudet ja mahdollistavat integroitumisen myös aikaisempien järjestelmien, kuten EDW:n (Enterprise Data Warehouse), RDBMS:än (Relational Database Management System) ja MPP:n kanssa. HDP integroituu ja laajentaa lisäksi jo käytössä olevia sovelluksia ja järjestelmiä. Tämän ansiosta saadaan Hadoopista hyöty mahdollisimman pienillä muutoksilla olemassa oleviin data-arkkitehtuureihin ja pienillä ammattitaidon lisäyksillä. Kaikki Hortonworksin tekemät ratkaisut ja kehitystyöt ovat edesauttamassa Apachen ohjelmistosäätiötä. (Hortonworks - Hortonworks Data Platform n.d.)

5.4.2 HDP:n versiot

HDP 2.4 on Hortonworks Sandbox

Suoritetaan virtuaalikoneella

- 32-bittinen ja 64-bittinen käyttöjärjestelmä (Windows 7, Windows 8 ja Mac OSX)
- Minimissään 10 gigatavua RAM-muistia
- Virtualisointi sallittuna BIOS:ista
- Selain: Chrome 25+, IE 9+, Safari 6+, Firefox 18+ suositus (Sandboxia ei voi suorittaa IE 10:llä)
- VirtualBox tai VMware Fusion (Hortonworks - Downloads n.d.)

HDP 2.4: Ready for the enterprise

1. Automatisoitu, sisältää Ambari 2.2:n. Suositeltu asennustapa HDP:n käyttöönottoon tuotantoympäristössä. Apache Ambari yksinkertaistaa klusterin resurssien varaamiset, hallinnan ja monitoroinnin. (Hortonworks - Downloads n.d.)

- 64-bittinen käyttöjärjestelmä tietyin versiorajoituksin (RHEL, CentOS, SLES, Ubuntu, Debian)

- Selain Firefox 18 tai Google Chrome 26
- RAM- ja ROM-muistien vaatimukset (ks. taulukko 1)

Taulukko 1. HDP 2.4 RAM- ja ROM-muistien vaatimukset

Isäntäkoneiden lkm	RAM	ROM
1	1024 MB	10 GB
10	1024 MB	20 GB
50	2048 MB	50 GB
100	4096 MB	100 GB
300	4096 MB	100 GB
500	8096 MB	200 GB
1000	12 288 MB	200 GB
2000	16 384 MB	500 GB

2. Manuaalinen (RPM, Red Hat Package Manager). Käytetään mikäli halutaan asentaa ja konfiguroida klusterit manuaalisesti RPM-paketeilla.

- 64-bittinen käyttöjärjestelmä tietyin versiorajoituksin (RHEL, CentOS, SLES, Ubuntu, Debian, Windows Server 2008 ja 2012)
- Ei yksittäisiä vaatimuksia laitteistosta asennuksen suhteen, mutta ROM-muistia on kuitenkin oltava minimissään 2,5 gigatavua. Laitteistovaatimukset määräytyvät täysin tulevien klustereiden mukaan. (Hortonworks - Downloads n.d.; Hortonworks - Ambari 2.2.1.1 n.d.)

Cloud:

Pilvipalvelu, sisältää Cloudbreak 1.2:n. Palvelua käytetään, mikäli halutaan pystyttää klusteri pilviympäristössä. Cloudbreak yksinkertaistaa klusterin resurssien varaamiset sekä hallinnan pilvessä. (Hortonworks - Downloads n.d.)

HDP-klustereita voidaan käyttää joko Cloudbreakin verkkokäyttöliittymällä tai CLI:llä (Command-Line Interface). Käytettävissä ovat julkiset pilvi-infrasturktuurisovellus-alustat, kuten Microsoft Azure, AWS ja GCP sekä yksityiset OpenStack-pilvi-infrastuk-tuurisovellusalustat. (Sequenceiq - Cloudbreak 1.2.0 n.d.)

Cloudbreakissa on kaksi pääkomponenttia, jotka ovat Cloudbreak Application ja Cloudbreak Deployer. Cloudbreak Application on tehty mikropalveluista, kuten Cloudbreak, Uluwatu sekä Sultans. Cloudbreak Deployer auttaa Cloudbreak Applicationin käyttöönotossa automaattisesti sisältäen Docker-tuen. Heti kun Cloudbreak Application on otettu käyttöön, sitä voidaan käyttää HDP-klustereiden kanssa eri pilviympäristöissä. Asennuksessa on kaksi eri vaihtoehtoa: (Mt.)

1. Cloudbreak Deployerin asentaminen omalle virtuaali- tai isäntäkoneelle.

- 64-bittinen käyttöjärjestelmä tietyin versiorajoituksin (RHEL, CentOS, Oracle Linux 7)
- Docker 1.9.1
- Vähintään 4 gigatavua RAM-muistia
- 10 gigatavua ROM-muistia
- Kaksiydinprosessori

2. Valmiiksi rakennettuja pilvi-imageja, joissa on Cloudbreak Deployer valmiiksi asennettuna

- Minimivaatimukset VM:lle
    - 4 gigatavua RAM-muistia
    - 10 gigatavua ROM-muistia
    - Kaksiydinprosessori
- Omat image-asennukset AWS:lle, GCP:lle sekä OpenStackille
- Azurelle ei ole valmista imagea, mutta optio Azure Resource Manager Templtesin käyttöön on olemassa, jonka kautta saa Cloudbreak Deployerin asennettua ja konfiguroitua. (Mt.)

Windows:

Ainoa julkaisu Hadoopista Windows-ohjelmistoalustalle.

- Minimissään 2,5 gigatavua ROM-muistia

- 64-bittinen käyttöjärjestelmä tietyin versiorajoituksin (Windows Server 2008 R2, 2012, 2012 R1 ja 2012 R2)
- Vaaditut ohjelmistot:
    - Python 2.7.x
    - Java JDK 1.7.x
    - Microsoft Visual C++ 2010 Redistributable Package (64-bit)
    - Microsoft .NET Framework 4.0 (Hortonworks - Downloads n.d.; Hortonworks - Windows Quick Start n.d.)

5.5 MapR

5.5.1 MapR Converged Data Platform

MapR Converged Data Platform integroi Hadoopin ja Sparksin, reaaliaikaiset tietokannat sekä maailmanlaajuisten tapahtumien suoratoistot yrityspuolen big data -tallennuksella kehittääkseen ja suorittaakseen innovatiivisia datasovelluksia. MapR-sovellusalusta toimii alan nopeiden, luotettavien, tietoturvallisten ja avoimien datainfrastruktuurien kanssa. Tämä vähentää huomattavasti kokonaiskustannuksia ja mahdollistaa maailmanlaajuiset reaaliaikaiset datasovellukset. MapR tukee kymmeniä avoimen lähdekoodin projekteja ja tarjoaa yritys-standardien mukaisen API-rajapinnan mahdollistaakseen työkalut, jotka asiakkaat tarvitsevat sovelluksilleen. (MapR - MapR Converged Data Platform 2016)

MapR Platform Services on joukko MapR Converged Data Platformin ydinosia. Näihin kuuluvat MapR Streams, MapR-DB ja MapR-FS. MapR Streams on maailmanlaajuinen tapahtumien julkaise-tilaa suoratoistojärjestelmä big datalle. MapR-DB on korkean suorituskyvyn omaava Hadoop NoSQL -tietokannan hallintajärjestelmä. MapR-FS on perustana oleva POSIX-tiedostojärjestelmä (Portable Operating System Interface for uniX), joka tarjoaa hajautetun, luotettavan, korkea suorituskykyisen, skaalautuvan sekä täyden luku- ja kirjoitusdatavaraston. (Mt.)

MapR tarjoaa vapaa valintaisia avoimen lähdekoodinratkaisuja tukemalla suosittuja avoimen lähdekoodin projekteja, jotka kehittävät datasovelluksia. Näitä ovat muun muassa Apache Hadoop, Apache Spark, Apache Drill ja Apache Search. (Mt.)

MapR:n avulla yritykset voivat hyödyntää jo investoituja käytössä olevia ohjelmistoja tai mitä tahansa työkaluja, jotka käyttävät standardin mukaista NFS-rajapintaa (Network File System). MapR tarjoaa yhtenäisen järjestelmän laskennalle, varastoinnille, verkolle sekä sovelluksille. (Mt.)

5.5.2 MapR Converged Data Platformin versiot

MapR 5.0:

MapR Converged Data Platformista on saatavilla kaksi eri versiota. MapR Converged Community Edition, joka on ilmainen sekä rajoittamaton tuotantokäyttöön tarkoitettu ratkaisu. MapR Converged Enterprise Edition (30 päivän kokeiluversio), joka on tarkoitettu kriittisiin yritysten käyttöönottoihin ja vaativat liiketoiminnan jatkuvuutta (HA/DR, High Availability/Disaster Recovery). (MapR - Distribution Editions 2016)

- 64-bittinen käyttöjärjestelmä tietyin versiorajoituksin (RHEL, CentOS, SUSE tai Ubuntu)
- Vähintään 8 gigatavua RAM-muistia (tuotannossa enemmän, tyypillisesti 32 GB)
- Vähintään 180 gigatavua ROM-muistia (Mt.)

MapR Sandbox:

Ilmainen MapR Sandbox on virtuaalikoneilla ajettava täysin ominaisuuksin varustettu klusteri. Mukana tulee kursseja, demosovelluksia sekä selainpohjainen käyttöliittymä, jotka mahdollistavat nopean tutustumisen ja aloituksen kehittäjille sekä ylläpitäjille. (MapR - Sandbox Hadoop 2016)

- VMware Player tai VirtualBox
- Vähintään 8 gigatavua RAM-muistia
- Vähintään 20 gigatavua ROM-muistia

- Neliydinprosessori
- Prosessori 64-bit tai x86 arkkitehtuuria
    - 1,3 GHz tai nopeampi AMD CPU (sekmenttirajoitus 64-bittisessä tilassa (long mode))
    - 1,3 GHz tai nopeampi Intel CPU (VT-x tuella) (Mt.)

MapR in the Cloud:

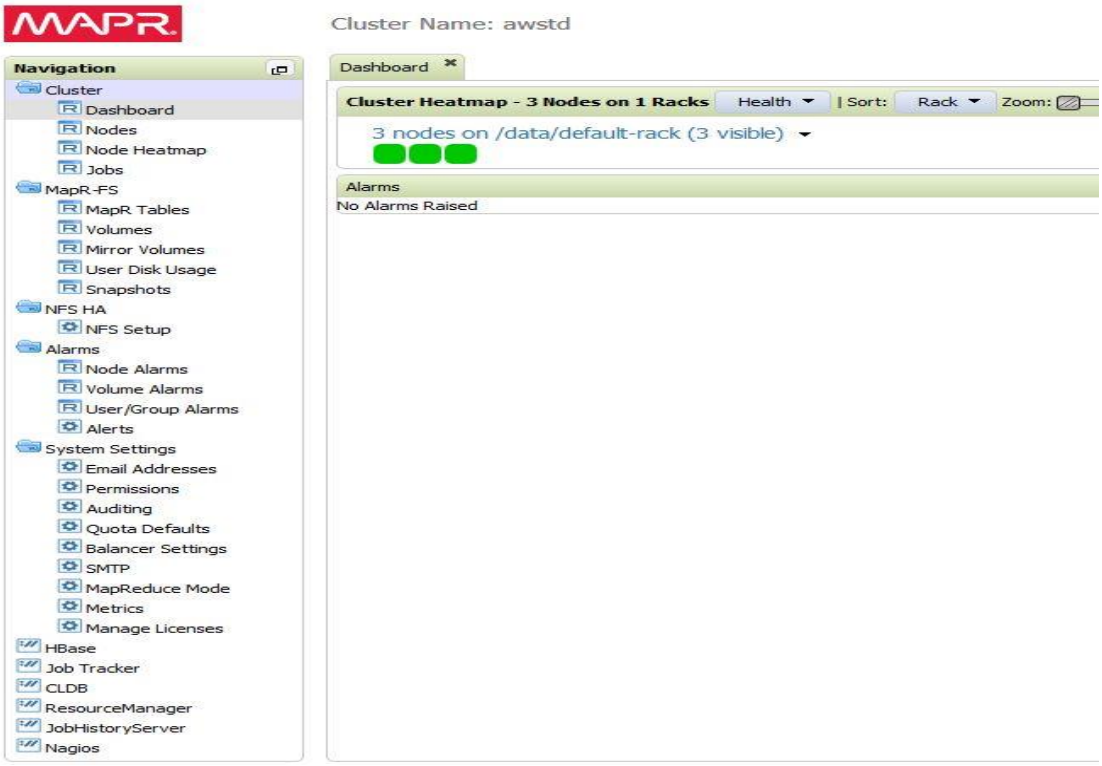
Mikäli käyttäjällä ei ole omia palvelimia, voi MapR:n ottaa käyttöön Azure-, GCP, AWS- tai CenturyLink-pilvipalveluiden kautta. Azure Fast Start sekä AWS Test Drive tarjoaa mahdollisuuden päästä tutustumaan tuotteisiin ilmaiseksi. (MapR - Hadoop as a Service 2016)

MapR in the Cloud -palvelu tarjoaa joustavan ja laajasti skaalautuvan pilvipalvelun Hadoopin käyttäjille. Metadata-arkkitehtuuri on täysin hajautettu POSIX-tiedostojärjestelmän mukaisesti. Korkean suorituskyvyn omaava järjestelmä sallii käyttäjien rakentaa yhtenäisiä klustereita, jotka ovat mahdollista skaalata tarpeiden mukaan helposti. Datan suodatukseen klustereissa käytetään standardien mukaisia käyttöliittymiä, jotka maksimoivat sijoitetun pääoman tuoton päivittäisistä tapahtumista. (Mt.)

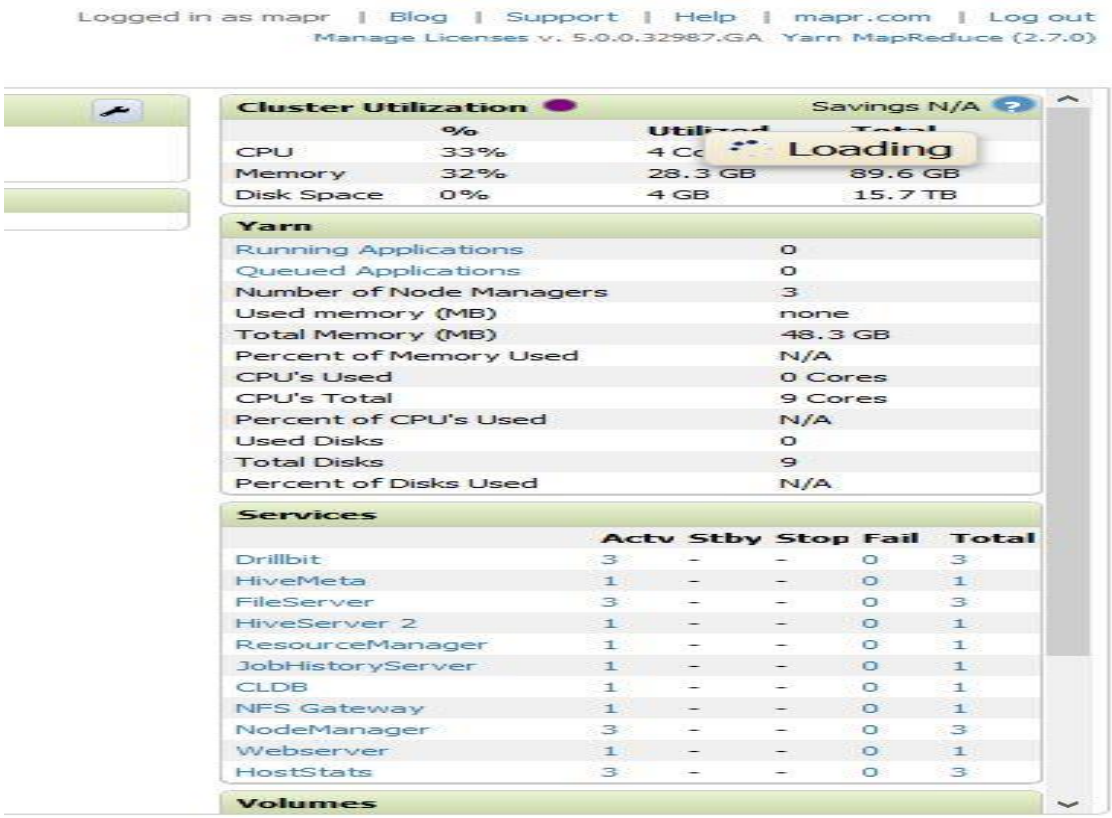
MapR Test Drive 5.0.0

MapR Test Drive for Hadoop on AWS on testiympäristö MapR:n pilvipalvelun kokeiluun kehittäjille, järjestelmän ylläpitäjille sekä yrityskäyttäjille (ks. kuvio 6 ja 7). Palvelu sisältää myös mahdollisuuden Apache Drillin käyttöön. Käyttöönottoa varten luovutetaan käyttäjän nimi-, osoite-, puhelinnumero- sekä sähköpostitiedot. Tilin varmistamisessa kestää noin kolme vuorokautta, jonka jälkeen sähköpostiin saa linkin ja ohjeet testiympäristön käytöstä. (MapR - Test Drive for Hadoop AWS 2016)

Palvelun käynnistys tapahtuu sähköpostiin saapuvan verkko-osoitteen kautta ja automattisesti tapahtuvien toimintojen eli tarvittavan instanssin luominen Amazon EC2 -palveluun sekä MapR-ohjelmiston asennus kestää noin 15 minuuttia. Tämän jälkeen käyttäjällä on kuusi tuntia aikaa testata palvelua sähköpostiin saapuvan uuden verkko-osoitteen, käyttäjätunnuksen ja salasanan kautta. Testaamista on tehty helpottamaan 11 sivuinen ohjeistus.



Kuvio 6. MapR-klusterin ohjauspaneeli osa 1 (Orbitera - Test Drives 2016)



Kuvio 7. MapR-klusterin ohjauspaneeli osa 2 (Orbitera - Test Drives 2016)

5.6 Jakeluiden vertailu

5.6.1 Yleistä

IBM, Cloudera, Hortonworks ja MapR ovat neljä johtavaa yritystä markkinoilla olevista Hadoop-jakeluista. Yrityspuolen Hadoop-jakeluilla on alle kymmenen vuotta vanhat markkinat, mutta arvion mukaan 100 % suuryrityksistä ottaa Hadoopin sekä siihen liittyvät teknologiat käyttöön seuraavan kahden vuoden sisällä. Markkinoiden suurimmat riskit ovat Clouderalla, Hortonworksilla ja MapR:llä, koska heidän tuotteensa ovat pelkästään Hadoop-ratkaisuja tai jollakin tavalla siihen liittyviä. Tällä hetkellä Hadoop-markkinoilla ei ole selvää johtajaa. Jokainen valmistaja keskittyy omiin avainominaisuuksiinsa, kuten tietoturvaan, skaalautuvuuteen, integroituvuuteen, hallintaan sekä suorituskykyyn. (Gualtieri ja Yuhanna 2016, 5)

Markkinajohtajien lisäksi Pivotal on varteenotettava yritys Hadoop-markkinoilla. Erityisesti tilanteessa, joissa yritykset näkevät saavansa hyötyä Pivotalin HAWQ SQL-for-Hadoop -moottorista sekä MADlib koneoppimisen kirjastosta. Lisäksi Pivotal on ODPi:n jäsen, joten osa Hadoop-jakelun komponenteista ovat vastaavia, kuin IBM:llä ja Hortonworksilla. (Gualtieri ja Yuhanna 2016, 6)

5.6.2 Forrester Waven pisteytykset

Yhdysvaltalainen tutkimuksia tekevä yritys Forrester Wave julkaisi tammikuussa 2016 viiden suurimman Hadoop-jakelun vertailun, jossa jaettiin pisteitä kolmessa eri kategoriassa: tämänhetkinen tarjonta, toimintasuunnitelma sekä markkinat (ks. kuvio 8). (Hadoop Market is Neck and Neck, Forrester Says 2016)

Tämänhetkisen tarjonnan -osiosta korkeimmat pisteet saavutti Cloudera ja alhaisimmat pisteet sai Pivotal. Cloudera pärjäsi erityisesti tietoturvassa, datassa sekä datan hallitsemisessa. MapR sai parhaat pisteet arkkitehtuurista, mutta alhaisimmat pisteet kehitystyöstä. Cloudera ja Hortonworks saivat parhaimmat pisteet ylläpidosta. (Mt.)

Hortonworks sai eniten pisteitä toimintasuunnitelma-osiosta. Forrester piti Hortonworksin ja MapR:n hankinta- ja hinnoittelustrategiaa muita kilpailijoita parempana. IBM oli ainoa, joka sai täydet pisteet sovellustuesta. Pivotalin pisteet olivat alhaisimmat toimintasuunnitelma-osiosta. (Mt.)

Markkina-osiosssa Clouderalla ja Hortonworksilla oli tasainen pisteytys muiden jäädessä jälkeen. (Mt.)

	Forrester's Weighting	Cloudera	Hortonworks	IBM	MapR Technologies	Pivotal Software
CURRENT OFFERING	50%	4.53	3.82	4.32	4.34	3.14
Solution configuration	5%	5.00	5.00	5.00	5.00	4.00
Architecture	20%	4.20	3.40	4.00	4.80	2.40
Administration	15%	5.00	4.75	3.75	4.25	3.75
Security	10%	5.00	3.00	4.32	4.34	3.00
Data	15%	4.25	3.50	3.50	4.75	3.00
Data governance	10%	5.00	3.00	5.00	3.00	3.00
Workload flexibility	10%	3.00	3.00	5.00	5.00	3.00
Development	10%	5.00	5.00	5.00	3.00	3.00
Platform integrations	5%	5.00	5.00	5.00	5.00	5.00
STRATEGY	50%	4.63	4.75	4.50	4.50	3.56
Acquisition and pricing	25%	4.50	5.00	3.00	5.00	2.25
Solution road map	25%	5.00	5.00	5.00	4.00	3.00
Ability to execute	25%	5.00	5.00	5.00	5.00	5.00
Implementation support	25%	4.00	4.00	5.00	4.00	4.00
MARKET PRESENCE	0%	4.56	4.45	3.33	3.78	2.21
Evaluated product revenue	33%	4.00	4.00	3.00	3.00	2.00
Customer base	33%	4.67	4.34	4.00	4.67	3.00
Partnerships	34%	5.00	5.00	3.00	3.66	1.66

All scores are based on a scale of 0 (weak) to 5 (strong).

Kuvio 8. Forrester Waven pisteytykset (Gualtieri ja Yuhanna 2016, 7)

5.6.3 Cloudera

Clouderan tuotteiden skaala ja kehitystahti ovat erottuvia muista kilpailijoista. Tämä ei kuitenkaan ole yllättävää, koska kyseessä on vuonna 2008 perustettu ensimmäinen kaupallinen Hadoop-yritys. Cloudera aloitti SQL-for-Hadoop villityksen Impalalla. Se tarjosi ensimmäisenä visuaalisen klusterin hallintatyökalun ja jatkaa merkittävällä ponnistelulla avainkomponenttien, kuten tietoturvan, korkean käytettävyyden sekä hallitsemisen ja ylläpidon eteen työskentelyllä. Cloudera tarjoaa strategiset hankinnat ja kumppanuuden yrityksille toimittaen samalla puuttuvat osat tietoturvasta, datan hallinnasta ja analysoinnista. Clouderan tärkeimmät kaupalliset arvot ovat

Cloudera Manager, Cloudera Navigator ja Impala -työkalut sekä yrityksen kokonaisvaltainen näkemys big data -sovellusalustasta. (Gualtieri ja Yuhanna 2016, 7)

Forrester ylisti Clouderaa trendien luonnista ja erityisesti viittasi Clouderan liikkeisiin SQL-on-Hadoop osa-alueella sekä lisäyksistä yritystason ominaisuuksiin, kuten tietoturvaan, korkeaan saatavuuteen ja hallintaan ratkaisuissaan. (Hadoop Market is Neck and Neck, Forrester Says 2016)

5.6.4 MapR

MapR:n tavoite alusta alkaen on ollut suunnitella jakelu, joka mahdollistaa Hadoopin täyden suorituskyvyn ja skaalautuvuuden potentiaalin mahdollisimman pienellä vaivalla. MapR on korvannut jakelussaan HDFS:än omalla MapRFS-tiedostojärjestelmällä. HDFS API:a käyttävä MapR:n täysi luku- ja kirjoitustiedostojärjestelmä MapRFS voi tallentaa biljoonia (10^{12}) tiedostoja, kun vaikeasti konfiguroitava HDFS vaatii erilliset nimiavaruudet. MapR on myös tehnyt enemmän töitä, kuin muiden jakeluiden tekijät, luotettavien ja tehokkaiden suurilukumääräisten klustereiden jakeluiden eteen. (Gualtieri ja Yuhanna 2016, 8; Cloudera vs Hortonworks vs MapR: Comparing Hadoop Distributions 2014)

Tyypillisesti MapR:n asiakkailla on jo olemassa tai he suunnittelevat isoja tehtäväkriittisiä Hadoop-klustereita ja haluavat käyttää MapR-DB:tä sekä MapR Streamsia. MapR:n ja Canonicalin yhteistyön ansiosta MapR M3 -julkaisu Hadoopista on tarjolla oletuskomponenttina Ubuntu-käyttöjärjestelmässä. MapR M3 on ilmainen, mutta siitä puuttuu joitakin komponentteja, kuten JobTracker HA (High Availability), NameNode HA, NFS-HA (Network File System) sekä Mirroring. (Mt.)

Forrester Waven mukaan MapR erottuu joukosta toimittamalla äärimmäisen suorituskyvyn sekä luotettavuuden skaalan. MapR sai erityismaininnan omasta tiedostojärjestelmästään. (Hadoop Market is Neck and Neck, Forrester Says 2016)

5.6.5 IBM

IBM integroi datanhallinnan komponentit sekä analyyttiset vahvuudet avoimen lähdekoodin ytimeen omassa Hadoop-jakelussaan. Yritykset, jotka jo käyttävät IBM:n datanhallintajärjestelmiä tuntevat luonnolliseksi lisätä BigInsights olemassa olevaan

datasovellusalustaansa. IBM on myös käynnistänyt kunniahimoisen avoimen lähdekoodin projektin Apache SystemML:än, joka on tarkoitettu Apache Sparkin koneoppimiselle. (Gualtieri ja Yuhanna 2016, 8)

IBM:n kaupalliset arvot tulevat Hadoop-laajennusten kehittyneisyydestä. Näistä esimerkkinä BigSQL, joka on yksi nopeimmista ja eniten SQL-taipuvainen kaikista SQL-for-Hadoop -koneista. Lisäksi BigQuality, BigIntegrate ja InfoSphere Big Match tarjoavat kehittyneet ja paljon ominaisuuksia sisältävät työkalut, jotka suorittavat natiivisti YARN:ia vaikeimmissakin Hadoopin käyttökohteissa. (Mt.)

5.6.6 Hortonworks

Hortonworksin vahvuus on 100 % avoimen lähdekoodin jakelu, ja kuinka se etsii muotoa kaikenkattavalle ja laajalle avoimen lähdekoodin yhteisölle. Kaikki HDP:ssä käytössä olevat teknologiat ovat Apachen avoimen lähdekoodin projekteja. Hortonworks vaatii yrityksiä paikkamaan aukkoja ja välittömästi julkaisevat koodit Apache-projektien käytettäviksi. Hortonworks ei myöskään aristele itse tehdä hankintoja täyttääkseen puuttuvia osia tuotevalikoimastaan ja sen asiakkaat pitävät avoimesta lähestymistavasta innovaatioihin. Lisäksi yritys on tärkeä jäsen ODPi:ssa. (Gualtieri ja Yuhanna 2016, 8; Hadoop Market is Neck and Neck, Forrester Says 2016)

5.6.7 Pivotal

Pivotal teki ison strategisen liikkeen, kun se päätti antaa oman osuutensa avoimen lähdekoodin yhteisölle. Pivotal jakoi yhteisölle käyttöön monet sen avainkomponenteista, kuten Greenplumin, GemFiren, HAWQ SQL-for-Hadoopin sekä MADlib koneoppimisen. Yrityksen päämääränä on tehdä Pivotal HD Hadoop -jakelusta isomman big data -sovellusalustan komponentti. Tällä pyritään parantamaan ja saamaan parempaan asemaan sen muita tietokanta-, datanhallinta- sekä sovellustuotteita. Pivotal on myös ODPi:n jäsen IBM:n ja Hortonworksin lailla. (Gualtieri ja Yuhanna 2016, 8)

5.6.8 Jakeluiden vertailu rinnakkain

Taulukossa 2 on vertailtu IBM:n, Clouderan, Pivotalin, Hortonworksin sekä MapR:n jakeluiden natiiviasennusten vaatimuksia RAM- ja ROM-muistien, prosessorien sekä

käyttöjärjestelmien suhteen. Ilmoitetut tiedot laitteistojen osalta ovat valmistajien ilmoittamia suositusvaatimuksia.

Taulukko 2. Natiiviasennusten vertailu

	IBM BigInsights for Apache Hadoop	**Cloudera Distribution Including Apache Hadoop (CDH)**	**Pivotal Big Data Suite**	**Hortonworks Data Platform (HDP)**	**MapR Converged Data Platform**
RAM (GB)	24	64	x	1	8
ROM (GB)	80	500	x	10	180
Prosessori	x	4-ydin	x	x	x
Käyttöjärjestelmä	x86 tai Power 64-bit Red Hat Linux	64-bittinen RHEL-compatible, CentOS, Oracle Linux, SUSE Linux, Ubuntu, Debian	x	64-bittinen RHEL, CentOS, SLES, Ubuntu, Debian, Windows Server 2008 ja 2012	64-bittinen RHEL, CentOS, SUSE tai Ubuntu

Taulukossa 3 on vertailtu virtualisointiratkaisuiden samoja vaatimuksia, kuin taulukossa kaksi. Näiden tietojen lisäksi taulukossa on virtualisointiohjelmistot, joilla jakelut toimivat.

Taulukko 3. Virtualisointiratkaisuiden vertailu

	IBM BigInsights for Apache Hadoop	**Cloudera Distribution Including Apache Hadoop (CDH)**	**Pivotal Big Data Suite**	**Hortonworks Data Platform (HDP)**	**MapR Converged Data Platform**
RAM (GB)	12	4 (CDH) 8 (Express) 10 (Enterprise)	x	10	8
ROM (GB)	50	x	x	x	20
Prosessori	4-ydin	x	x	Moniydin	4-ydin
Käyttöjärjestelmä	x	x	x	32- ja 64-bittinen käyttöjärjestelmä (Windows 7, Windows 8 ja Mac OSX)	x
Virtualisointiohjelmistot	VMware Windows tai OS X	VMware, KVM ja VirtualBox	x	VMware Fusion ja VirtualBox	VMware Player tai VirtualBox

Taulukossa 4 on vertailtu pilvipalveluratkaisuiden vaatimuksia, kuten taulukoissa kaksi ja kolme. Taulukoon neljä on lisätty hintavertailu sekä pilvipalveluratkaisuiden käyttämät integraatioalustat.

Taulukko 4. Pilvipalveluratkaisuiden vertailu

	IBM BigInsights on Cloud	**Cloudera Director**	**Pivotal Cloud Foundry (PCF)**	**Hortonworks Cloudbreak**	**MapR in the Cloud**
RAM (GB)	x	4	x	4	x
ROM (GB)	x	8	x	10	x
Prosessori	x	2-ydin	x	2-ydin	x
Käyttöjärjestelmä	x	64-bittinen RHEL ja CentOS 6.5, 6.7 ja 7.1, Ubuntu 14.04	x	64-bittinen RHEL, CentOS, Oracle Linux 7	x
Hinta	x	Perus (ilmainen) Kehittäjä (44€/kk) Yritys (alkaen 89€/kk)	x	x	x
Integraatioalustat	IBM Bluemix	AWS ja GCP	x	AWS, GCP ja OpenStack	AWS, Azure, GCP ja CenturyLink

Taulukossa 5 on vertailtu jakeluiden natiiviasennusten käytössä olevat teknologiat, ohjelmistot ja laitteistot.

Taulukko 5. Käytössä olevat teknologiat, ohjelmistot ja laitteistot

	IBM BigInsights for Apache Hadoop	**Cloudera Distribution Including Apache Hadoop (CDH)**	**Pivotal Big Data Suite**	**Hortonworks Data Platform (HDP)**	**MapR Converged Data Platform**
Analyyttinen tietokannanhallinta	DB2 ja Netezza	HBase ja lisäksi Cloudera Impala tukee SQL tietokantakyselyjä Hadoopin päällä.	Pivotal Greenplum	HBase ja lisäksi Hortonworks Hive tukee SQL tietokanta-kyselyjä Hadoopin päällä.	HBase ja lisäksi Drill, Hive, Impala, Shark sekä muita SQL tietokantakyselyjä Hadoopin päällä tukevia ratkaisuja.
Sisäisen muistin tietokannanhallinta	DB2 with BLU Acceleration ja solidDB	Apache Spark tukee sisäisen muistin analysointia Hadoopin päällä.	Pivotal GemFire ja SQLFire. Pivotal HD käyttää yhdistelmää GemFire XD:stä ja HAWQ:sta sisäisen	Apache Spark tukee sisäisen muistin analysointia Hadoopin päällä.	MapR mainostaa sisäisen muistin suorituskykyä avoimen lähdekoodin projekteilla

			muistin analysointiin Hadoopin päällä.		Drill ja Shark.
Suoratoiston prosessointitekнologia	InfoSphere Streams	Avoimen lähdekoodin suoratoistoprosessoinnin vaihtoehdot, mukaan lukien Storm.	Pivotalilla on käynnissä projekti, jossa tavoitteena on GemFiren ja SQLFiren integrointi Pivotal Hadoopin ja Spring XD:n kanssa datan suodatusmekanismiksi, joka tukisi skaalautuvaa suoratoistoanalyysia.	Avoimen lähdekoodin suoratoistoprosessoinnin vaihtoehdot, mukaan lukien Storm.	MapR tukee suoratoistoanalysointia Stormilla sekä Informatican ja HParserin integraatiolla.
Laitteisto ja ohjelmisto	PureData System For Operational Analytics (DB2), IBM PureData System for	Yhteistyökumppanien Cisco, Dell, HP, IBM, NetApp, ja	Pivotal Data Computing Appliance	Yhteistyökumppanien HP, Teradata sekä muiden (val-	Laitteisto konfiguraatiot mahdollisia Ciscolta, HP:ltä, IBM:ltä

	Analytics (Netezza); PureData System for Hadoop (BigInsights).	Oracle (valmiiksi konfiguroidut) laitteet.		miiksi konfiguroidut) laitteet.	sekä NetAppilta.

Natiiviasennusten suuri skaala laitteistojen muistien osalta johtuu siitä, että Hortonworks ilmoittaa suositusvaatimuksensa yhden tietokoneen mukaan. Klustereiden ollessa samankokoisia eivät vaatimukset eroa suuresti toisistaan. Selkeän eron natiiviasennuksissa tekee Hortonworks, jonka jakelu on mahdollista asentaa ainoana myös Windows Server 2008 ja 2012 käyttöjärjestelmille tietyin versiorajoituksin.

Virtualisointiratkaisuissa ei vaatimusten osalta ole merkittäviä eroja. Jakeluiden toimimattomuus jollakin tietyllä virtualisointiohjelmistolla voi olla rajoittava tekijä jakelun valintaa tehtäessä.

Pilvipalveluratkaisuiden laitteistovaatimuksissa ei ole merkittäviä eroja, koska pilvipalveluissa kuormitus tapahtuu ulkoisilla palvelimilla. Valintaa tehtäessä jakeluiden integraatio mahdollisuudet eri alustoille kannattaa ottaa huomioon myös tulevaisuutta ajatellen. Jakeluista MapR:llä on tällä hetkellä laajimmat integrointiominaisuudet.

Vaikka laitteistovaatimusten erot kaikkien jakeluiden kesken ovat pieniä, on hyvä huomioida, että suurissa satojen tai tuhansien tietokoneiden klustereissa voi näistäkin pienistä eroista kasvaa iso kustannuserä.

Suurimmat erot jakeluiden vertailuissa tulevat teknologioissa, ohjelmistoissa ja laitteistoissa. Valintaa tehtäessä pitää tutkia myös nykyisiä jo käytössä olevia ratkaisuja ja niiden yhteensopivuutta siirtyessä big datan hallintaan.

5.6.9 Kustannukset

Big Data tuotteiden hintoja ei ole suoraan nähtävillä valmistajien kotisivuilla, joten suoraa hintavertailua on mahdoton tehdä. Lisäksi täysin samanlaisia ratkaisuja sa-

moilla ominaisuuksilla ei ole kaikilta yrityksiltä saatavissa. Cloudera ilmoittaa poikkeuksellisesti Directorin kuukausihinnoittelut, mutta hintaan ei kuitenkaan sisälly AWS:n ja GCP:n maksut, joiden päällä sovellusta ajetaan.

Todellinen hinnoittelu esimerkki löytyy Oraclen Big Data -ryhmältä, joka sisältää esiasennetun ja konfiguroidun järjestelmän Cloudera CDH:lla ja sen kaikilla mahdollisilla optioilla varustettuna (Oracle - Price Comparison for Big Data Appliance 2014):

- Big Data laitteisto (mukana automaattinen tukipyyntö komponenttien vikatilanteista).
- Cloudera CDH ja Cloudera Manager.
- Kaikki Clouderan optiot sekä Accumulo ja Spark.
- Oracle Linux ja Oracle JDK.
- Oracle R -jakelu.
- Oracle NoSQL Database Community Edition.
- Oracle Big Data Appliance Enterprise Manager Plug-In.

Oraclen järjestelmien Premier-tuki on kiinteä hintainen 55 859€ vuodessa. (Mt.)

Taulukossa 6 on kuvattu kolmen vuoden kulut kyseiselle ratkaisulle. (Mt.)

Taulukko 6. Oraclen Big Data kustannukset

	1. vuosi	**2. vuosi**	**3. vuosi**	**3 vuotta yhteensä**
Big Data laitteiston kustannukset	465 489€	-	-	465 489€
Vuosittainen tukipalvelu	55 859€	55 859€	55 859€	167 577€
On-site asennus	12 413€	-	-	12 413€
Yhteensä	533 761€	55 859€	55 859€	645 479€

645 479 eurolla saa 18 kpl Sun X4-2L palvelinta, 288 ydintä (kaksi Intel Xeon E5-2650V2 prosessoria jokaisessa tietokoneessa), 864 teratavua kovalevytilaa (12 kappaletta 4 teratavun kovalevyä jokaisessa tietokoneessa), tarvittavat ohjelmistot, tukipalvelun sekä asennuksen paikan päällä ja konfiguroinnin. Yhdelle teratavulle tulee hintaa kyseisessä ratkaisussa 747 euroa kolmessa vuodessa. (Mt.)

Vertailuna pilvipalvelun hinnoittelusta IBM:n Bluemix-alustalla toimivan BigInsights for Apache Hadoopin 18 palvelimen klusteri maksaa noin 42 230€/kk pelkällä avoimen lähdekoodin Apache Hadoop -jakelulla. Paketti sisältää kuitenkin tallennustilaa 1728 teratavua, joka on yli puolet enemmän, kuin Oraclen tarjoamassa ratkaisussa. Tässä ratkaisussa yhdelle teratavulle tulee hinnaksi 880€ kolmessa vuodessa. Bluemix-alustalla halvin klusteri (3 laskentakonetta ja 48 teratavua tallennustilaa) on hinnalta noin 11 200€/kk. (Bluemix - Define Cluster 2016)

6 Testaus

6.1 IBM Bluemix

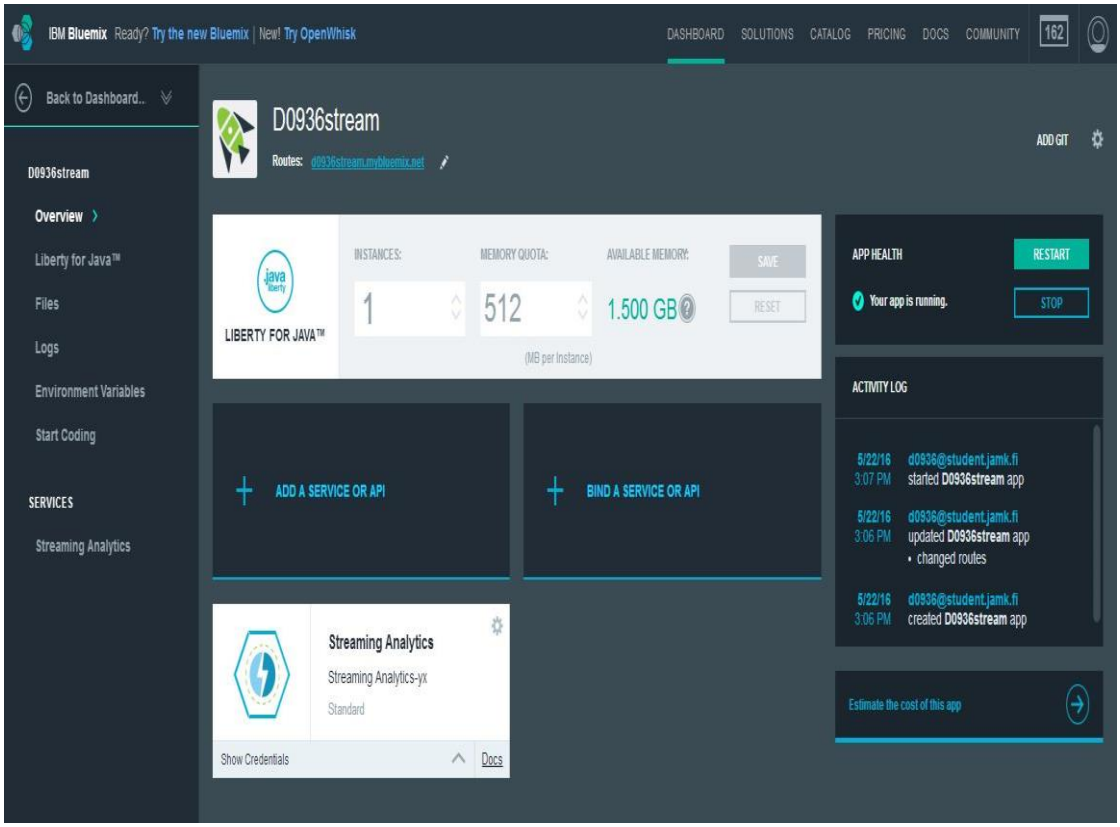
Big data -jakeluista testaukseen valikoitui IBM:n Bluemix-alusta ja sen ominaisuudet big datan suhteen. IBM:ltä saadut opiskelijatunnukset mahdollistivat riittävän pitkän ajan tutustua Bluemixiin, mutta oikeiden klustereiden korkean hinnan vuoksi varsinaista BigInsights for Apache Hadoopia ei voitu testata. Näin ollen työssä keskityttiin kolmeen muuhun big data -ominaisuuteen. Tarkasteluun otettiin Streaming Analytics, Insights for Twitter sekä Insights for Weather -palvelut. Dataa ja sen analysointia varten IBM tarjoaa 17 erilaista palvelua (ks. kuvio 9) sekä neljä kolmannen osapuolen palvelua. Kolmannen osapuolen palvelut ovat ClearDB MySQL Database, ElephantSQL, Namara.io Catalog ja Redis Cloud.

Kuvio 9. IBM Bluemixin data- ja analysointipalvelut (Bluemix - Catalog 2016)

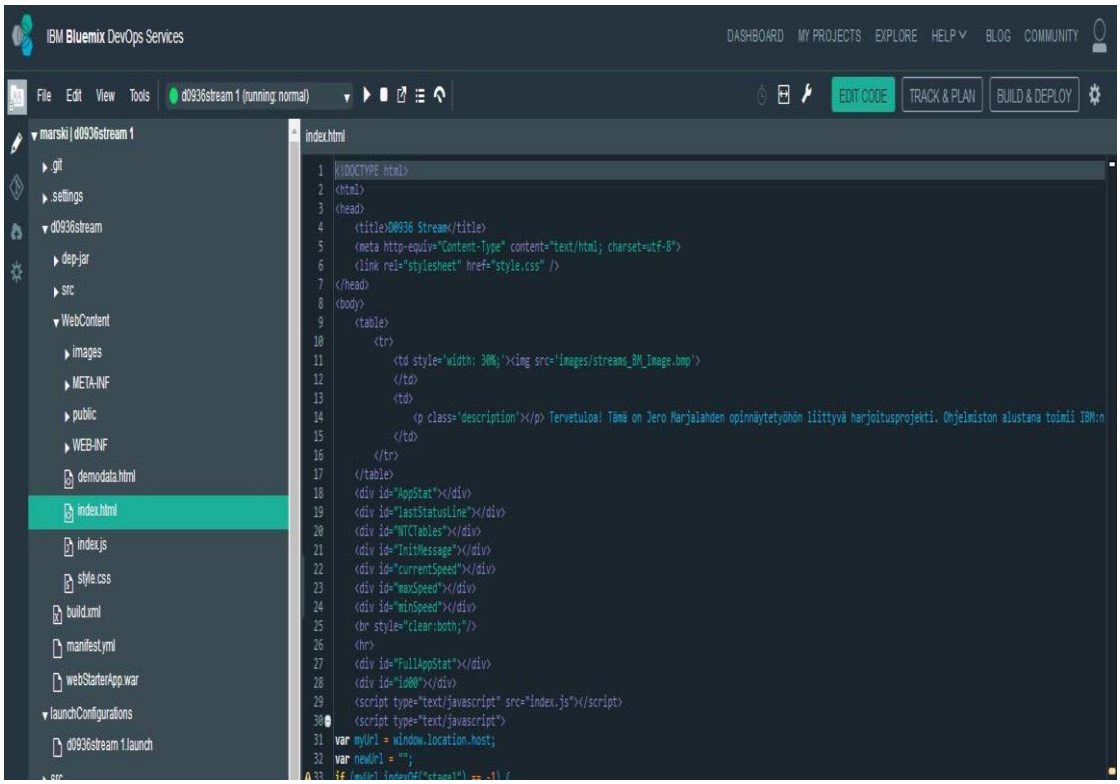
6.2 Streaming Analytics

Streaming Analytics on tehty reaaliaikaisen datavirran suodattamiseen, analysointiin sekä monitorointiin. Kyseinen palvelu lisätään joko omaan tai valmiiseen Bluemix sovellukseen ja se mahdollistaa miljoonien tapahtumien analysoinnin sekunnissa. Streaming Analyticsia voi testata 120 tuntia ilmaiseksi. (Bluemix - Streamin Analytics 2016)

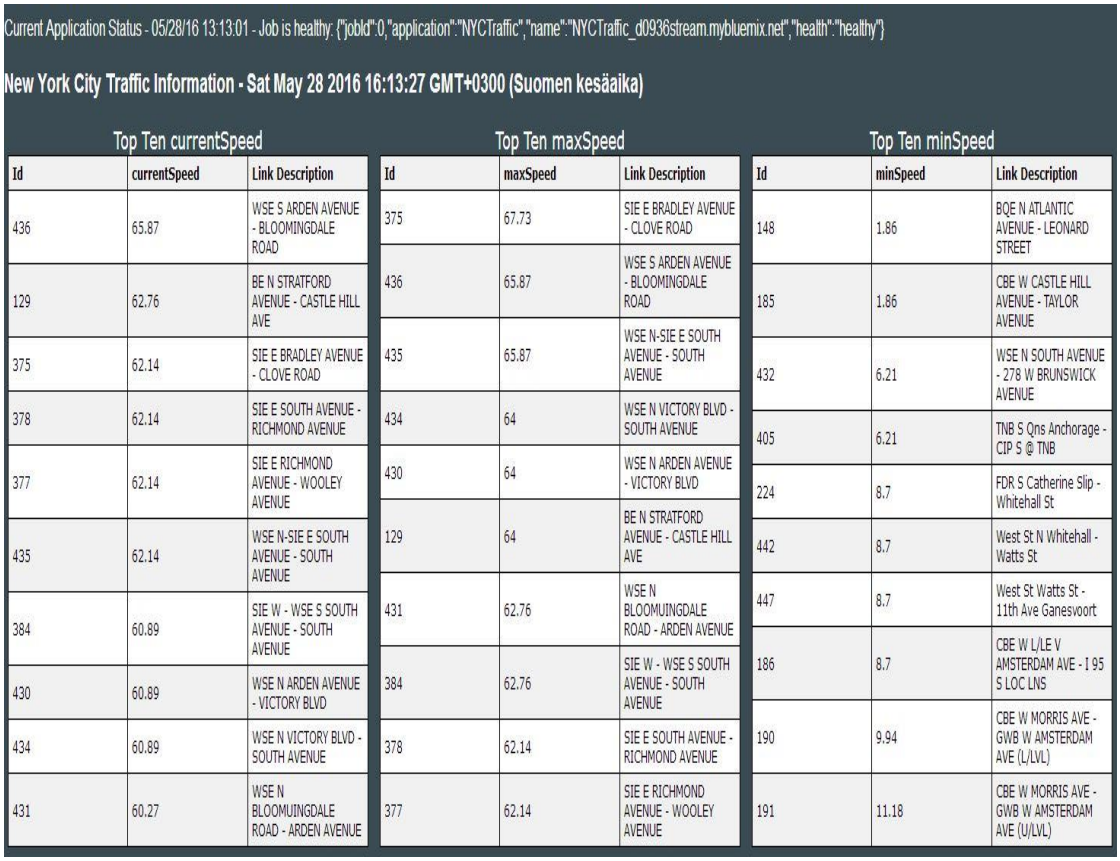
Testauksessa otettiin käyttöön omalle Bluemix-alustalle (ks. kuvio 10) IBM:n valmis sovellus, joka on toteutettu Liberty for Javalla. Valmiin koodin muokkauksessa, koonti- sekä käyttöönottovaiheessa käytettiin CLI:n sijasta Bluemixin omaa DevOps Services -palvelua (ks. kuvio 11). Tämä palvelu mahdollistaa testauksen täysin pilviympäristössä, ilman mitään omalle tietokoneelle asennettavia ohjelmia tai ajureita. Lisäksi DevOps mahdollistaa eri projektien hallinnan samalla käyttöliittymällä. Sovelluksella tarkkaillaan New York Cityn liikenteen ajoneuvojen nopeuksia (ks. kuvio 12). Data ohjelmaan tulee New York City Department of Transportationin (DOT) julkiselta verkkosivustolta. Tarkka kuvaus testiohjelman käyttöönotosta vaihe vaiheelta löytyy liitteestä 1.



Kuvio 10. IBM Bluemix -sovelluksen ohjauspaneeli (Bluemix - Application Dashboard 2016)



Kuvio 11. IBM Bluemix DevOps Services (Hub Jazz Net - IBM Bluemix DevOps Services 2016)



Current Application Status - 05/28/16 13:13:01 - Job is healthy: {"jobId":0,"application":"NYCTraffic","name":"NYCTraffic_d0936stream.mybluemix.net","health":"healthy"}

New York City Traffic Information - Sat May 28 2016 16:13:27 GMT+0300 (Suomen kesäaika)

Top Ten currentSpeed

Id	currentSpeed	Link Description
436	65.87	WSE S ARDEN AVENUE - BLOOMINGDALE ROAD
129	62.76	BE N STRATFORD AVENUE - CASTLE HILL AVE
375	62.14	SIE E BRADLEY AVENUE - CLOVE ROAD
378	62.14	SIE E SOUTH AVENUE - RICHMOND AVENUE
377	62.14	SIE E RICHMOND AVENUE - WOOLEY AVENUE
435	62.14	WSE N-SIE E SOUTH AVENUE - SOUTH AVENUE
384	60.89	SIE W - WSE S SOUTH AVENUE - SOUTH AVENUE
430	60.89	WSE N ARDEN AVENUE - VICTORY BLVD
434	60.89	WSE N VICTORY BLVD - SOUTH AVENUE
431	60.27	WSE N BLOOMUINGDALE ROAD - ARDEN AVENUE

Top Ten maxSpeed

Id	maxSpeed	Link Description
375	67.73	SIE E BRADLEY AVENUE - CLOVE ROAD
436	65.87	WSE S ARDEN AVENUE - BLOOMINGDALE ROAD
435	65.87	WSE N-SIE E SOUTH AVENUE - SOUTH AVENUE
434	64	WSE N VICTORY BLVD - SOUTH AVENUE
430	64	WSE N ARDEN AVENUE - VICTORY BLVD
129	64	BE N STRATFORD AVENUE - CASTLE HILL AVE
431	62.76	WSE N BLOOMUINGDALE ROAD - ARDEN AVENUE
384	62.76	SIE W - WSE S SOUTH AVENUE - SOUTH AVENUE
378	62.14	SIE E SOUTH AVENUE - RICHMOND AVENUE
377	62.14	SIE E RICHMOND AVENUE - WOOLEY AVENUE

Top Ten minSpeed

Id	minSpeed	Link Description
148	1.86	BQE N ATLANTIC AVENUE - LEONARD STREET
185	1.86	CBE W CASTLE HILL AVENUE - TAYLOR AVENUE
432	6.21	WSE N SOUTH AVENUE - 278 W BRUNSWICK AVENUE
405	6.21	TNB S Qns Anchorage - CIP S @ TNB
224	8.7	FDR S Catherine Slip - Whitehall St
442	8.7	West St N Whitehall - Watts St
447	8.7	West St Watts St - 11th Ave Ganesvoort
186	8.7	CBE W L/LE V AMSTERDAM AVE - I 95 S LOC LNS
190	9.94	CBE W MORRIS AVE - GWB W AMSTERDAM AVE (L/LVL)
191	11.18	CBE W MORRIS AVE - GWB W AMSTERDAM AVE (U/LVL)

Kuvio 12. New York Cityn liikenneinformaatio (Mybluemix - New York City Traffic Information 2016)

6.3 Insights for Twitter

IBM Insights for Twitter tarjoaa reaaliaikaisen prosessoinnin yhteisö- ja mikroblogipalvelu Twitterin datavirrasta. Käytössä on laaja valikoima konfiguroitavia hakuparametreja sekä avainsanoja. Palvelu sisältää REST (REpresentational State Transfer) ohjelmointirajapinnan, joka mahdollistaa hakujen kustomoinnin ja palauttaa ”tweetit” eli viestit JSON-formaatissa (JavaScript Object Notation). Twitter-sisältöä voi etsiä Twitter Decahosesta, joka sisältää satunnaisesti valitut 10 % kaikista Twitter-viesteistä tai maksullisesta PowerTrack-suoratoistosta, joka sisältää koko Twitterin indeksoidun sisällön. Ilmaiseksi viestejä voi hakea viisi miljoona kappaletta. Määrä lasketaan hakutuloksista palautuvien viestien lukumäärästä. (Bluemix - Insights for Twitter 2016)

Testauksessa käytettiin samoja valintoja kuin Streaming Analyticsissä, eli toteutus on tehty Liberty for Javalla, ja koodin muokkauksessa, koonti- sekä käyttöönottovai-

heessa käytettiin DevOps Services -palvelua. Käyttöön otettiin IBM:n Twitter-hakupalvelu. Hakupalvelun avulla pystytään hakemaan pelkkä hakusanojen lukumäärä Twitteristä tai myös viestit, jotka sisältävät kyseisen hakusanan (ks. kuvio 13). Avaamalla ”Insights”-linkin viestin alta päästään näkemään kyseisen viestin rikastettu data (ks. kuvio 14). Käyttämällä luonnollista kieliprosessointia Insights for Twitter analysoi viestit ja päättelee, onko viestin sisältö sävyltään positiivinen, neutraali vaiko negatiivinen. Testatessa huomattiin, että analysointi ei toimi suomenkielisistä viesteistä. Esimerkki viestissä näkyy analysoinnin tuloksena positiivinen sävy kohdassa ”Content”, johon on päädytty sanasta ”Great”. Twitter-linkin avaamalla päästään kyseisen viestin metadataan (ks. kuvio 15). Tarkka kuvaus testiohjelman käyttöönotosta vaihe vaiheelta löytyy liitteestä 2.

Kuvio 13. Twitter-hakutulos sanalla #JAMK (Mybluemix - IBM Insights for Twitter 2016)

IBM **Insights** for Twitter

cde	author	gender	unknown		
		parenthood	isParent	unknown	
			evidence		
		location	country	Russia	
			city		
			state		
		maritalStatus	isMarried	unknown	
			evidence		
	content	sentiment	evidence	polarity	POSITIVE
				sentimentTerm	Great
			polarity	POSITIVE	

Kuvio 14. Twitter-viestin sävyn analysointi (Mybluemix - IBM Insights for Twitter 2016)

JAMK @JAMK_fi

"Älkää rakastuko ensimmäiseen ideaan", kuuluu ohje innovaatioviikolla. #jamk #innovaatioviikko #designthinking http://t.co/cDXufP9NsB

IBM Insights for **Twitter**

message	twitter_extended_entities	media	display_url	pic.twitter.com/cDXufP9NsB		
			indices	111 133		
			sizes	small	w	340
					h	190
					resize	fit
				large	w	1024
					h	574
					resize	fit
				thumb	w	150
					h	150
					resize	crop
				medium	w	600
					h	336
					resize	fit
			id_str	574917123732029440		
			expanded_url	http://twitter.com/JAMK_fi/status/574917124583518208/photo/1		
			media_url_https	https://pbs.twimg.com/media/B_qEE_dWQAAcgYe.jpg		
			id	574917123732029440		
			type	photo		
			media_url	http://pbs.twimg.com/media/B_qEE_dWQAAcgYe.jpg		
			url	http://t.co/cDXufP9NsB		
	postedTime	2015-03-09T12:58:18.000Z				
	verb	post				
	link	http://twitter.com/JAMK_fi/statuses/574917124583518208				
	generator	displayName	Twitter Web Client			
		link	http://twitter.com			
	body	"Älkää rakastuko ensimmäiseen ideaan", kuuluu ohje innovaatioviikolla. #jamk #innovaatioviikko #designthinking http://t.co/cDXufP9NsB				
	favoritesCount	0				
	objectType	activity				
	actor	summary	JAMK University of Applied Sciences is an attractive, internationally oriented higher education institution in Jyväskylä, Finland.			
		image	jamk.fi			
		statusesCount	830			
		utcOffset	7200			
		languages	en			
		preferredUsername	JAMK_fi			

Kuvio 15. JAMK:in Twitter-viestin metadata (Mybluemix - IBM Insights for Twitter 2016)

6.4 Insights for Weather

IBM Insights for Weather palvelu antaa integroida säähistoriaa ja reaaliaikaisia säätietoja The Weather Companyn datasta Bluemix-sovelluksiin. Käytössä on tämän hetkinen sää, yhden vuorokauden tunnin välinen ennuste, 10 päivän ennuste sekä 24 tuntia vanhat säätiedot. Datan avulla voidaan esimerkiksi ennustaa, havaita ja visualisoida tämänhetkisiä ja tulevia sääilmiöitä. Säähakuja voi tehdä ilmaiseksi 10 hakua tunnissa ja 500 hakua vuorokaudessa. (Bluemix - Insights for Weather 2016)

Sääsovelluksen toteutus on tehty SDK for Node.js:llä, mutta muuten käyttöönotto tapahtui samanlailla, kun Streaming Analyticsin ja Insights for Twitterin testauksessa. Sääsovellus näyttää ennusteet vuorokauden jokaiselta tunnilta, 10 päivän ennusteet sekä menneet havainnot 24 tunnin ajalta (ks. kuvio 16). Saatavat tiedot ovat myös graafisennäkymän lisäksi saatavilla JSON-formaatissa (ks. kuvio 17). Tarkka kuvaus testiohjelman käyttöönotosta vaihe vaiheelta löytyy liitteestä 3. Opinnäytetyön aikana Insights for Weather korvattiin Weather Company Data for IBM Bluemix nimisellä palvelulla, joten liitteessä on ohjeet korvaavalla versiolla tehtynä.

Kuvio 16. Insights for Weatherin graafinen-näkymä (Mybluemix - Insights Weather 2016)

Kuvio 17. Insights for Weatherin data JSON-formaatissa (Mybluemix - Insights Weather 2016)

7 Yhteenveto

Opinnäytetyön tavoitteena oli JYVSECTEC:in toimesta tutustua yleisellä tasolla big dataan sekä tutkia tarkemmin ennalta valittujen viiden suurimman big data -toimittajan ratkaisut.

Kirjallisuutta ja verkkomateriaalia big data aiheesta löytyi paljon, mutta tietojen nopea vanhentuminen sekä puuttuvat yhtenevät termit vaikeuttivat kokonaisuuden hahmottamista sekä kirjoittamista selkeäksi kokonaisuudeksi. Tästä huolimatta big

datasta yleisesti saatiin selkeä kuva rakennettua opinnäytetyöhön ja tutustuminen meni jopa syvällisemmäksi teorian osalta.

Big data jakeluiden vertailu tapahtui vain teoriatasolla, eikä todellisia ominaisuuksia ja toimintoja päästy vertailemaan käytännössä. Toimittajien kotisivujen mainoslauseet saatiin karsittua pois, kun vertailussa ja testauksessa esiin nousseita ongelmia alkoi esiintyä. Vertailusta saatiin selville suurimpien big data -toimittajien eri vaihtoehdot, vahvuudet ja heikkoudet sekä ohjelmisto- ja laitteistovaatimukset.

Testausosiossa ongelmaksi nousivat käytettävissä olevat resurssit. Pilvipalveluiden todelliset kustannukset sekä natiivi- ja virtuaaliasennusten suuret laitteistovaatimukset estivät useimpien jakeluiden testauksen kotoa käsin. Testaukseen olisi ollut mahdollista ottaa mukaan koulun testiympäristöön VPN-yhteydellä (Virtual Private Network) toteutettu virtuaaliklusteri, mutta tästä luovuttiin opinnäytetyön edetessä pilvipalveluiden suuntaan.

Toimittajat ilmoittivat pilvipalveluiden ilmaisista kokeilujaksoista, mutta todellisuudessa ne olivat lähinnä vain tutustumisia käyttöliittymiin. Toinen ongelma pilvipalveluiden testaamisessa oli, että jakelut itsessään olivat ilmaisia tai niissä oli ilmaiset kokeilujaksot, mutta alustat joilla jakeluita ajettiin, kuten Amazon Web Services ja Google Cloud Platform, olivat tuntihinnoiteltuja. Tosin jakeluiden toimittajilta ei voida vaatia suuria klustereita ja korkeaa laskentatehoa pelkille kokeilujaksoille ilman korvausta. Todennäköisesti todellisessa ostotilanteessa mahdollisuudet testauksiin ovat monipuolisempia ja lisäksi silloin on myös tarkka kuva siitä, mitä tarvitaan, joten hintavertailu on myös helpompaa.

Näiden ongelmien kautta päädyttiin tutkimaan syvällisemmin IBM Bluemixin tarjoamia pilvipalvelun mahdollisuuksia. Päätöstä tuki myös se, että IBM:ltä saatiin opiskelijatunnukset pidempiaikaiseen käyttöön. IBM:n klustereiden hinnoittelu sulki pois mahdollisuuden testata BigInsights for Apache Hadoopia pilvipalvelussa. Bluemixin data- ja analyysipalveluista löytyi kuitenkin hyviä vaihtoehtoja, joista valikoitui kolme testaukseen. Näitä sai testata ilmaiseksi tietyin rajoituksin. Näillä kolmella palvelulla saatiin aikaiseksi testiohjelmat, joilla pystyttiin testaamaan käytännössä Bluemixin tarjoamia mahdollisuuksia eri tilanteissa.

Työn tavoitteet saavutettiin suppeaksi jäänyttä testausta lukuun ottamatta hyvin. Big datasta ja siihen liittyvistä ratkaisuista sekä niiden hyödyntämisestä saatiin laaja kokonaisvaltainen kuva. Bluemixin tarjoamista mahdollisuuksista löytyi myös hyviä harjoituksia Jyväskylän ammattikorkeakoulussa tulevaisuudessa järjestettävää big data -koulutusta ajatellen.

Lähteet

AWS Amazon - Cloud Formation. 2016. Console.aws.amazon.com. Viitattu 27.3.2016. Https://us-west-2.console.aws.amazon.com/console/home

AWS Amazon - Console Home. 2016. Console.aws.amazon.com. Viitattu 27.3.2016. Https://eu-central-1.console.aws.amazon.com/cloudformation/designer/home

AWS Amazon - Getting Started. N.d. Aws.amazon.com. Viitattu 26.3.2016. Http://aws.amazon.com/getting-started/

AWS Amazon - Quick Start. N.d. Aws.amazon.com. Viitattu 26.3.2016. Https://aws.amazon.com/quickstart/

AWS Amazon - Sign Up. N.d. Portal.aws.amazon.com. Viitattu 26.3.2016. Portal.aws.amazon.com/billing/signup

Bigdata - Big Data -määritelmiä. N.d. Bigdata.fi. Viitattu 23.1.2016. Http://bigdata.fi/big-data-maaritelma

Bluemix - Analytics for Apache Hadoop. 2015. Bluemix.net. Viitattu 2.3.2016. Https://console.ng.bluemix.net/catalog/services/analytics-for-apache-hadoop

Bluemix - Application Dashboard. 2016. Bluemix.net. Viitattu 28.5.2016. Https://console.ng.bluemix.net/?direct=classic/#/resources/orgGuid=93b96af2-55b2-472d-8c7a-af911f9e0d04&spaceGuid=d0ca8193-63a7-426f-88f9-6708121f294f&appGuid=f9d3403b-aff8-446c-adf5-370de87305f4&detailType=overview&paneId=2

Bluemix - BigInsights for Apache Hadoop. 2015. Bluemix.net. Viitattu 13.3.2016. Https://console.ng.bluemix.net/catalog/services/biginsights-for-apache-hadoop

Bluemix - Catalog. 2016. Bluemix.net. Viitattu 28.5.2016. Https://console.ng.bluemix.net/catalog/

Bluemix - Dashboard. 2016. Bluemix.net. Viitattu 27.3.2016. Https://console.eu-gb.bluemix.net/?/&direct=classic/#/resources/

Bluemix - Define Cluster. 2016. Bluemix.net. Viitattu 28.5.2016. Https://biginsightsoncloudv2.services.dal.bluemix.net:8443/clusters/new

Bluemix - Insights for Twitter. 2016. Bluemix.net. Viitattu 28.5.2016 Https://console.ng.bluemix.net/catalog/services/insights-for-twitter

Bluemix - Insights for Weather. 2016. Bluemix.net. Viitattu 28.5.2016. Https://console.ng.bluemix.net/catalog/services/insights-for-weather/

Bluemix - Streamin Analytics. 2016. Bluemix.net. Viitattu 28.5.2016. Https://console.ng.bluemix.net/catalog/services/streaming-analytics/

Cloudera - Apache Hadoop. N.d. Cloudera.com. Viitattu 19.3.2016. Https://cloudera.com/products/apache-hadoop.html

Cloudera - Director 2.0. N.d. Cloudera.com. Viitattu 26.3.2016. Http://www.cloudera.com/downloads/director/2-0-0.html#

Cloudera - Director. N.d. Cloudera.com. Viitattu 26.3.2016. Http://www.cloudera.com/products/cloudera-director.html

Cloudera - Key CDH Components. N.d. Cloudera.com. Viitattu 19.3.2016. Https://cloudera.com/products/apache-hadoop/key-cdh-components.html

Cloudera - Predictive and Proactive Support. N.d. Cloudera.com. Viitattu 19.3.2016. Https://cloudera.com/services-support/support-offering/predictive-and-proactive-support.html

Cloudera vs Hortonworks vs MapR: Comparing Hadoop Distributions. 2014. Experfy.com. Viitattu 30.4.2016. Https://www.experfy.com/blog/cloudera-vs-hortonworks-comparing-hadoop-distributions/

Cloudtweaks - Surprising Facts and Stats About The Big Data Industry. 2015. Cloudweaks.com. Viitattu 23.1.2016. Http://cloudtweaks.com/2015/03/surprising-facts-and-stats-about-the-big-data-industry/

EMC - DSSD D5. 2016. Emc.com. Viitattu 3.4.2016. Http://www.emc.com/en-us/storage/flash/dssd/dssd-d5/index.htm

Forbes - How Real-Time Weather Data Is Helping Businesses Run Better. 2015. Forbes.com. Viitattu 23.1.2016. Http://www.forbes.com/sites/ibm/2015/08/27/how-businesses-everywhere-are-benefitting-from-a-deluge-of-weather-data-photos/#7986d920547d

Gualtieri, M. & Yuhanna, N. 2016. Cloudera.com. The Forrester Wave™: Big Data Hadoop Distributions, Q1 2016. Viitattu 24.4.2016. Https://www.cloudera.com/content/dam/www/static/documents/analyst-reports/forrester-wave-big-data-hadoop-distributions.pdf

Hadoop Market is Neck and Neck, Forrester Says. 2016. Datanami.com. Viitattu 30.4.2016. Http://www.datanami.com/2016/01/20/hadoop-market-is-neck-and-neck-forrester-says/

Hortonworks - Ambari 2.2.1.1. N.d. Hortonworks.com. Viitattu 27.3.2016. Http://docs.hortonworks.com/HDPDocuments/Ambari-2.2.1.1/bk_Installing_HDP_AMB/content/_operating_systems_requirements.html

Hortonworks - Downloads. N.d. Hortonworks.com. Viitattu 27.3.2016. Http://hortonworks.com/hdp/downloads/

Hortonworks - Hortonworks Data Platform. N.d. Hortonworks.com. Viitattu 27.3.2016. Http://hortonworks.com/hdp/

Hortonworks - Windows Quick Start. N.d. Hortonworks.com. Viitattu 28.3.2016. Http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.4.0-Win/bk_QuickStart_HDPWin/content/index.html

Hub Jazz Net - IBM Bluemix DevOps Services. 2016. Hub.jazz.net. Viitattu 28.5.2016. Https://hub.jazz.net/code/edit/edit.html#/code/file/marski-OrionContent/marski%2520%257C%2520d0936stream%25201/d0936stream/WebContent/index.html

Hurwitz, J., Nugent, A., Halper, F. & Kaufman, M. 2013. Big Data For Dummies.

IBM - BigInsights for Apache Hadoop. N.d. 03.ibm.com. Viitattu 5.3.2016. Http://www-03.ibm.com/software/products/en/ibm-biginsights-for-apache-hadoop

IBM - BigInsights Quick Start. N.d. Ibm.com. Viitattu 2.3.2016. Http://www.ibm.com/analytics/us/en/technology/hadoop/

Jyvsectec - Tietoa meistä. 2016. Jyvsectec.fi. Viitattu 28.1.2016. Http://jyvsectec.fi/fi/tietoa-meista/

MapR - Converged Data Platform. 2016. Mapr.com. Viitattu 28.3.2016. Https://www.mapr.com/products/mapr-converged-data-platform

MapR - Distribution Editions. 2016. Mapr.com. Viitattu 28.3.2016. Https://www.mapr.com/products/mapr-distribution-editions

MapR - Hadoop as a Service. 2016. Mapr.com. Viitattu 28.3.2016. Https://www.mapr.com/products/hadoop-as-a-service

MapR - Sandbox Hadoop. 2016. Mapr.com. Viitattu 28.3.2016. Https://www.mapr.com/products/mapr-sandbox-hadoop/download

MapR - Test Drive for Hadoop AWS. 2016. Mapr.com. Viitattu 3.4.2016. Https://www.mapr.com/mapr-test-drive-for-hadoop-aws

Mybluemix - IBM Insights for Twitter Demo App. 2016. Mybluemix.net. Viitattu 28.5.2016. Https://d0936-tweet.mybluemix.net/

Mybluemix - Insights Weather. 2016. Mybluemix.net. Viitattu 28.5.2016. Http://insights-weather-d0936.mybluemix.net/

Mybluemix - New York City Traffic Information. 2016. Mybluemix.net. Viitattu 28.5.2016. Https://d0936stream.mybluemix.net/

Oracle - Price Comparison for Big Data Appliance. 2014. Oracle.com. Viitattu 17.4.2016 Https://blogs.oracle.com/datawarehousing/entry/updated_price_comparison_for_big

Orbitera - Test Drives. 2016. Mapr.com. Viitattu 3.4.2016. Https://mapr.orbitera.com/c2m/customer/testDrives/view/640

Pivotal - Big Data Suite. N.d. Pivotal.io. Viitattu 20.3.2016. Http://pivotal.io/big-data/pivotal-big-data-suite

Pivotal - Gemfire. N.d. Pivotal.io. Viitattu 20.3.2016. Http://pivotal.io/big-data/pivotal-gemfire

Pivotal - Getting Started With Pivotal Cloud Foundry. N.d. Pivotal.io. Viitattu 27.3.2016.
Http://pivotal.io/platform/pcf-tutorials/getting-started-with-pivotal-cloud-foundry/introduction

Pivotal - Greenplum. N.d. Pivotal.io. Viitattu 20.3.2016.
Http://pivotal.io/big-data/pivotal-greenplum

Pivotal - HDB. N.d. Pivotal.io. Viitattu 20.3.2016.
Http://pivotal.io/big-data/pivotal-hdb

Pivotal - Madlib. N.d. Pivotal.io. Viitattu 27.3.2016.
Http://pivotal.io/madlib

Pivotal - Web Services Dashboard. 2016. Pivotal.io. Viitattu 27.3.2016.
Https://console.run.pivotal.io/organizations

Salo, I. 2013. Big Data Tiedon vallankumous.

Salo, I. 2014. Big data & pilvipalvelut.

Sawant, N. & Shah, H. 2013. Big Data Application Architecture Q & A: A Problem-Solution Approach

Sequenceiq - Cloudbreak 1.2.0. N.d. Sequenceiq.com. Viitattu 28.3.2016.
Http://sequenceiq.com/cloudbreak-docs/release-1.2.0/

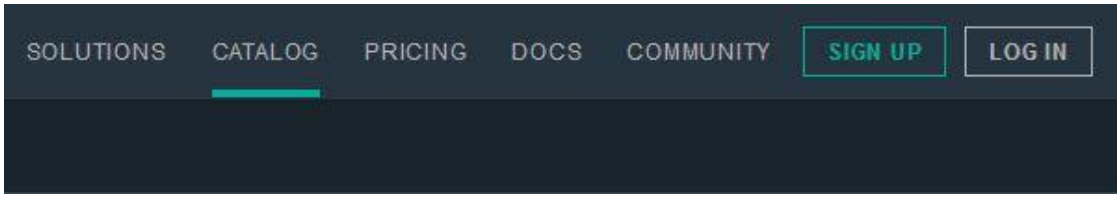
Spring - Project Spring XD. N.d. Spring.io. Viitattu 20.3.2016.
Http://projects.spring.io/spring-xd/

Liitteet

Liite 1. Streaming Analyticsin käyttöönotto

1. Tilien rekisteröinti

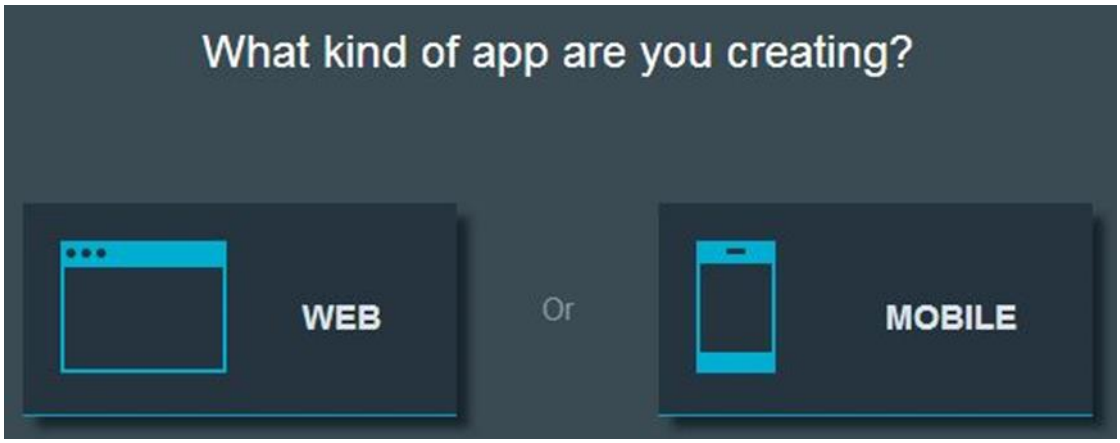
- Rekisteröi Bluemix-tili osoitteessa https://console.ng.bluemix.net/registration. Tämän jälkeen ota käyttöön DevOps-tili osoitteessa https://hub.jazz.net. Tilit ovat linkitetty yhteen henkilökohtaisen IBM ID:n kautta.



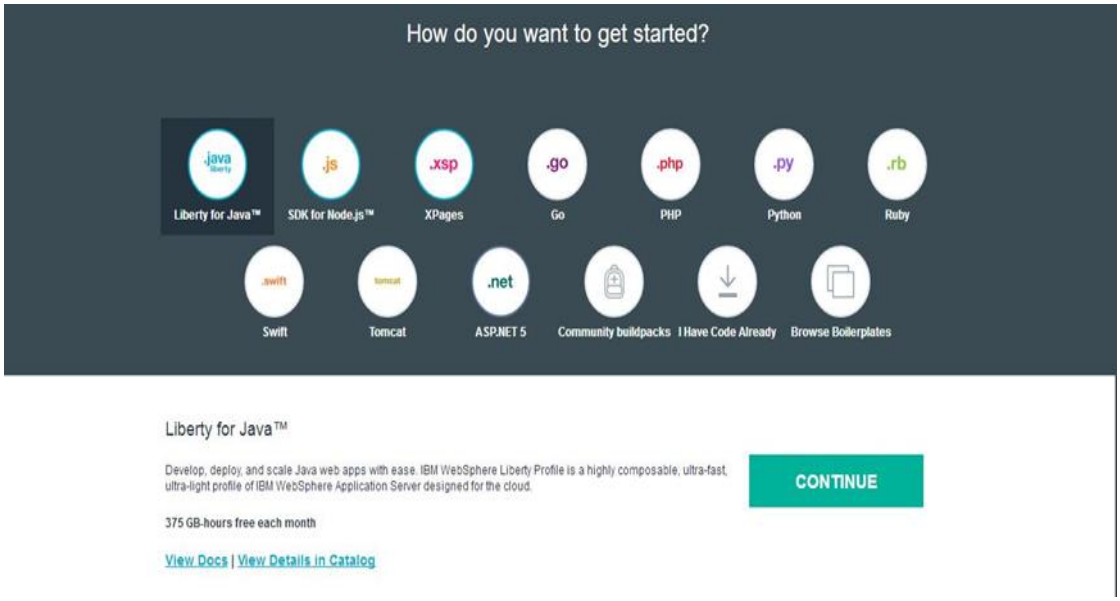
2. Bluemix sovelluksen luominen ja Streaming Analyticsin yhdistäminen siihen

- Kirjaudu Bluemix-palveluun osoitteessa https://console.ng.bluemix.net ja lisää sovellus valitsemalla ”Create an App”.

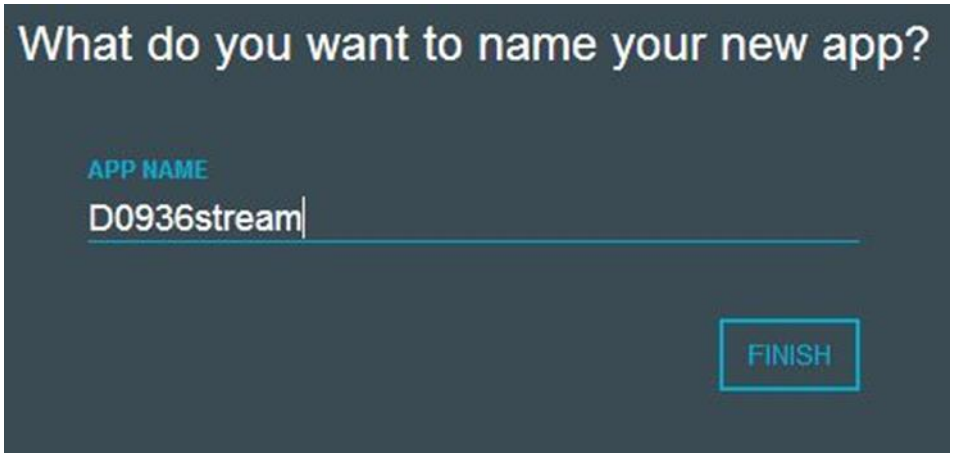
- Valitse sovelluksen tyypiksi ”Web”.



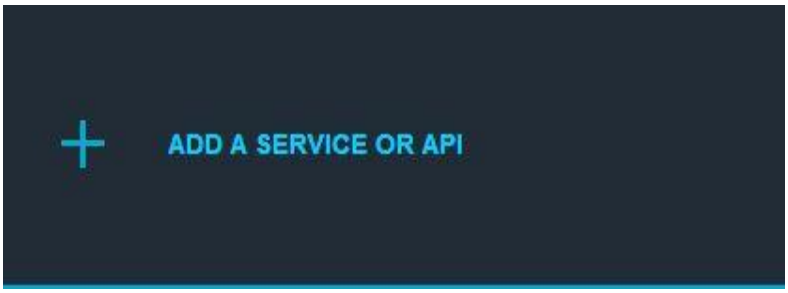
- Valitse käytettävissä olevista vaihtoehdoista ”Liberty for Java”.

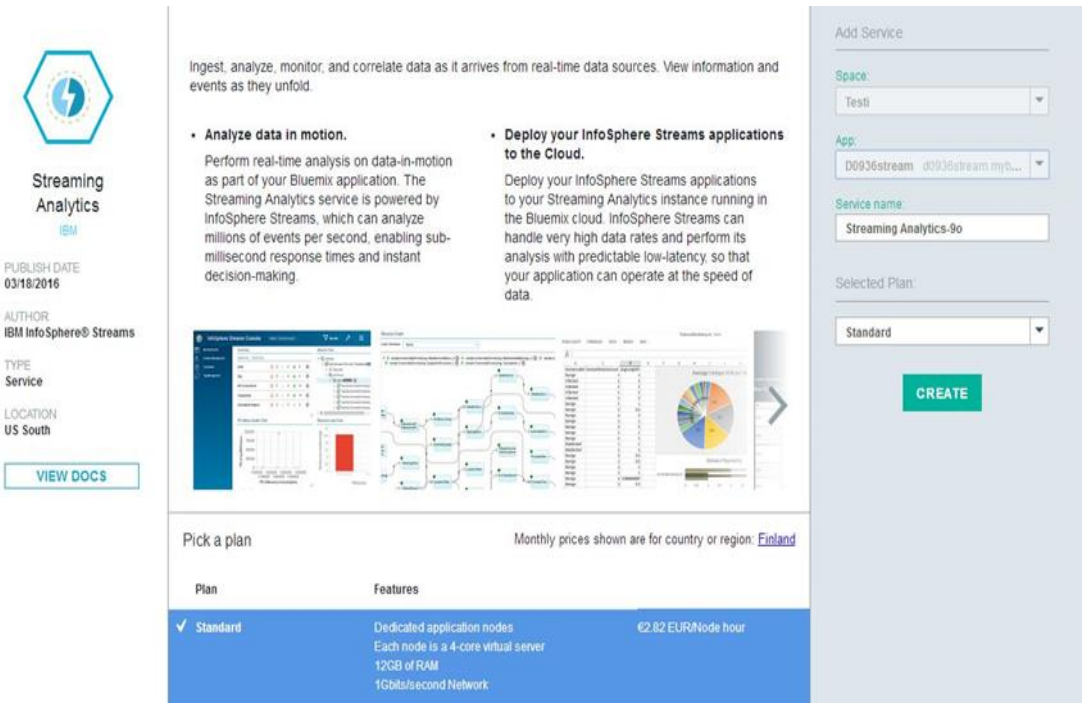


- Anna uudelle sovellukselle nimi.

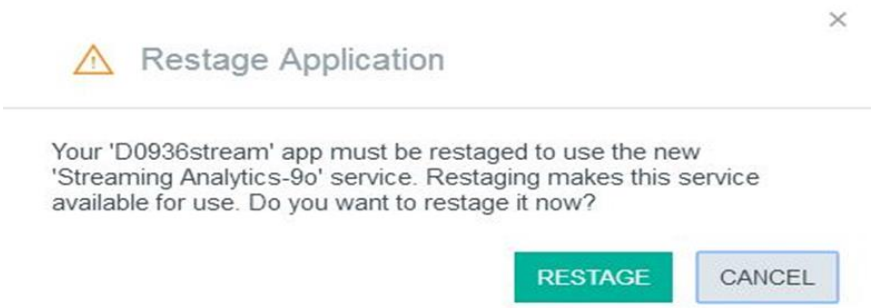


- Luodun ohjelman ”Overview” -sivulta valitse ”Add a Service or API”. Valitse palveluista Streaming Analytics ja valitse ”Create”.





- Valitse ”Restage” ilmoituksen tullessa näkyviin.



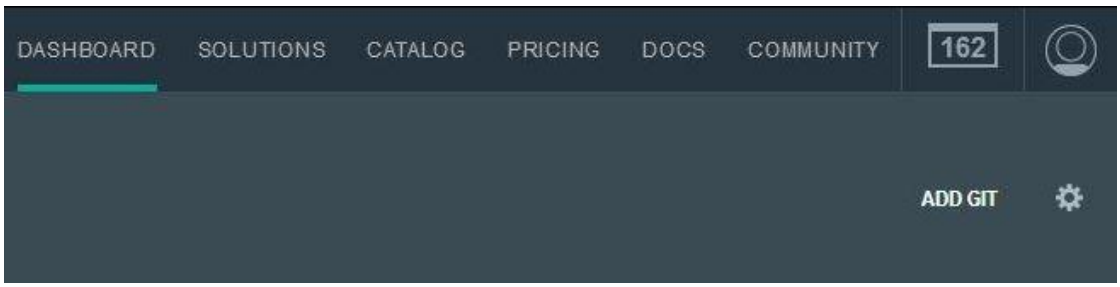
3. Lähdekoodin hankinta

- Lataa lähdekoodin .zip-tiedosto osoitteesta https://hub.jazz.net/git/streams-cloud/NYCTraffic/archive?revstr=master.

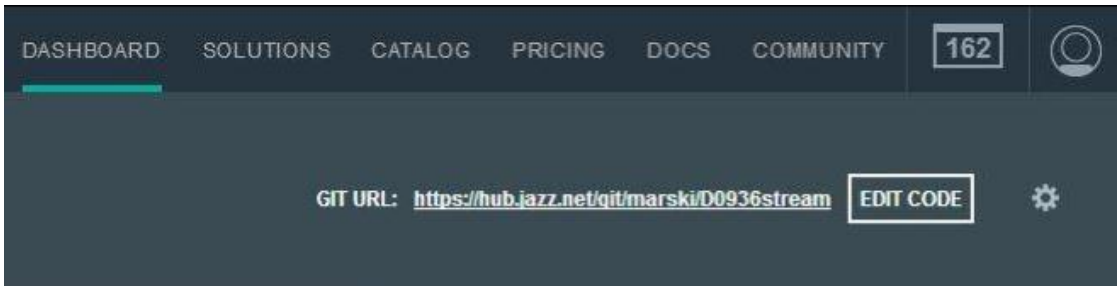
- Vaihtoehtoisesti voit myös ”forkata” alkuperäisen projektin tai kloonata git-tietolähteen. Alkuperäinen projekti löytyy DevOpsin valinnasta ”Explore” hakusanalla "streamscloud". Projektin nimi on “streamscloud | NYCTraffic”.

4. Sovelluksen käyttöönotto

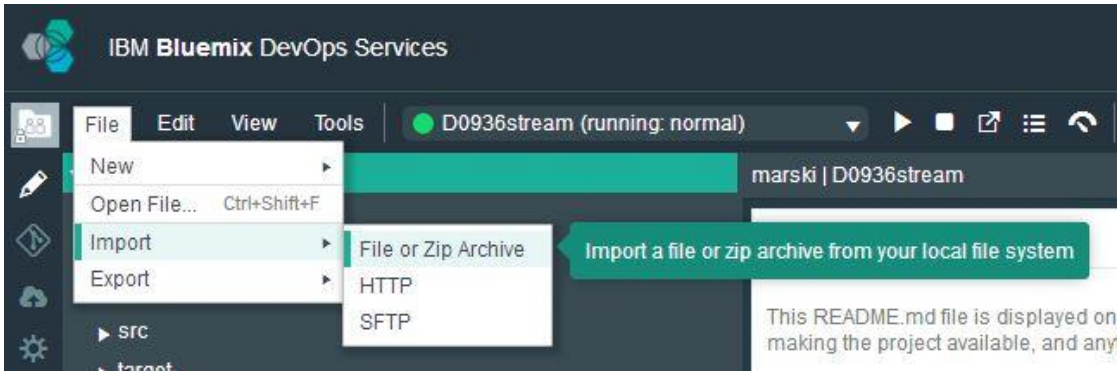
- Sovelluksen ”Overview” -sivulta valitse oikealta ylhäältä ”Add Git” ja valitse ”Continue”.



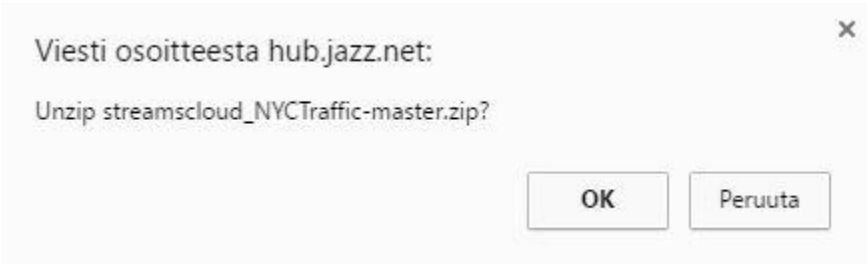
- Sovelluksen ”Overview” -sivulta valitse oikealta ylhäältä ”Edit Code”

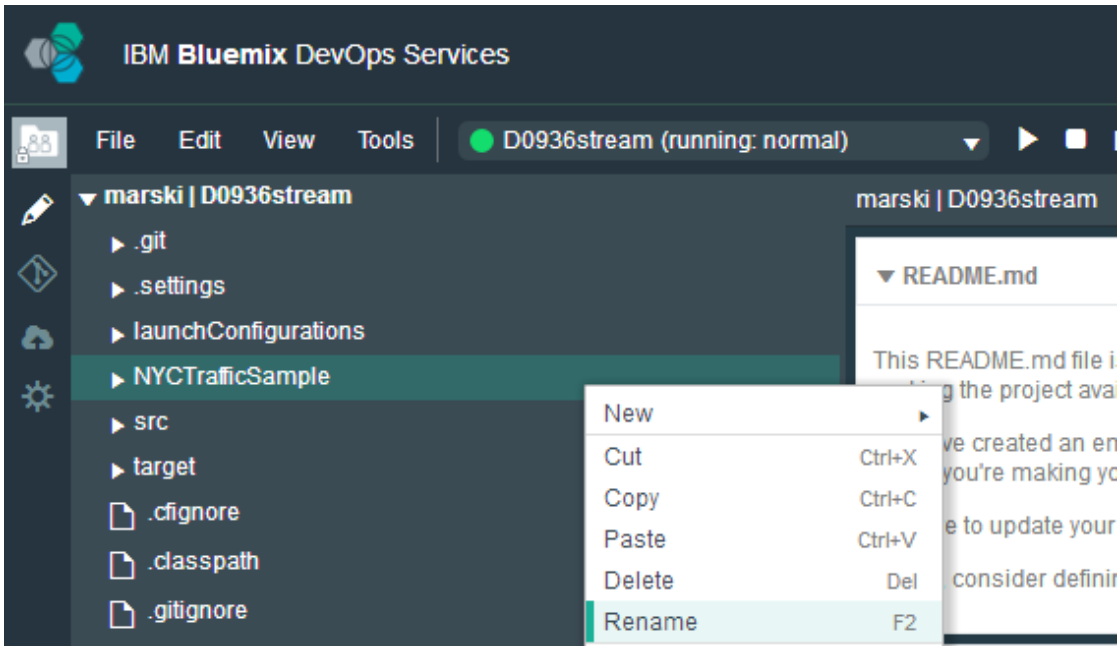


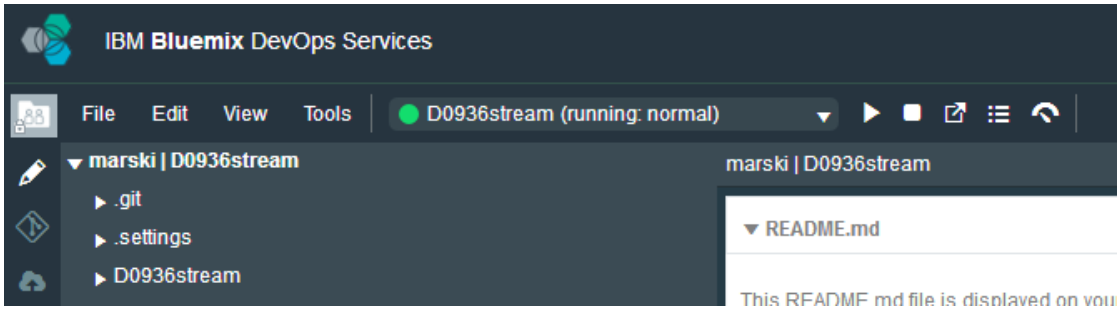
- Tuo lataamasi lähdekoodi projektiin valitsemalla ”File - Import - File or ZIP Archive”.



- Anna DevOpsin purkaa .zip-tiedosto latauksen yhteydessä ja nimeä purettu kansio ”NYCTrafficSample” kohdassa kaksi annetun uuden sovelluksen nimen mukaisesti.







- Muokkaa manifest.yml tiedostoon path-riville purkamasi kansion nimi/webStarterApp.war.

manifest.yml

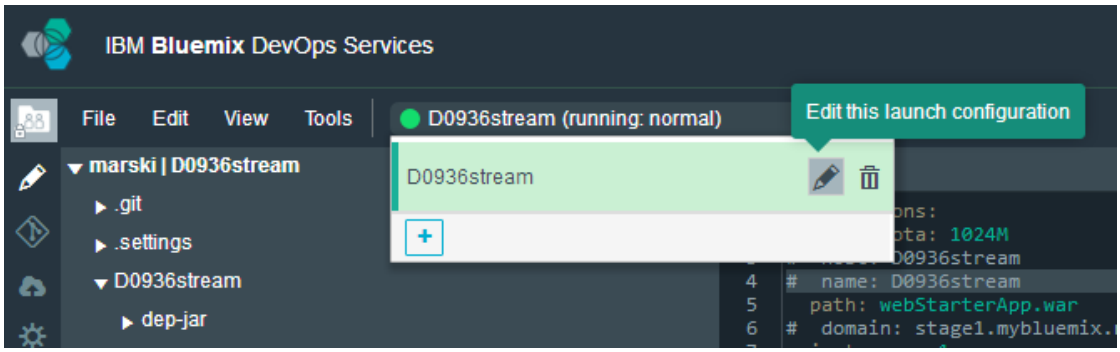
```
applications:
- path: D0936stream/webStarterApp.war
  memory: 512M
  instances: 1
  domain: mybluemix.net
  name: D0936stream
  host: d0936stream
  disk_quota: 1024M
  services:
  - Streaming Analytics-9o
```

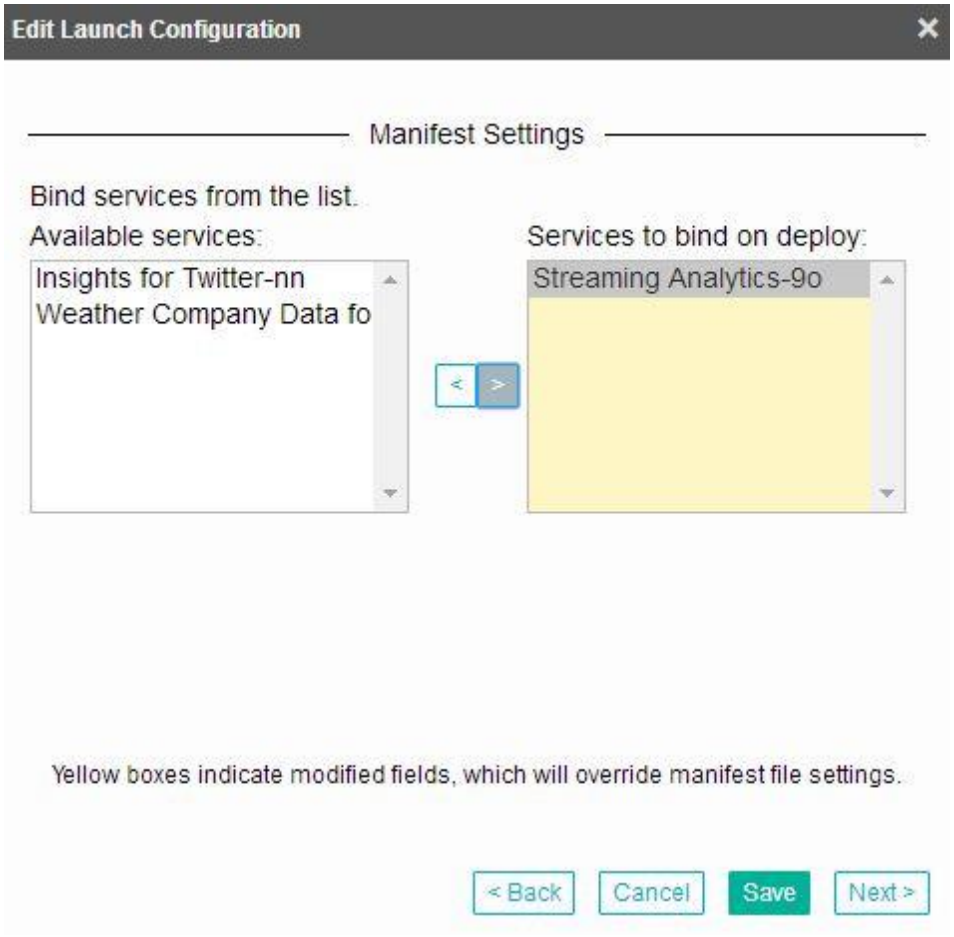
- Muokkaa luomasi kansion manifest.yml tiedoston host- ja name-riveille sovelluksesi nimi.

manifest.yml

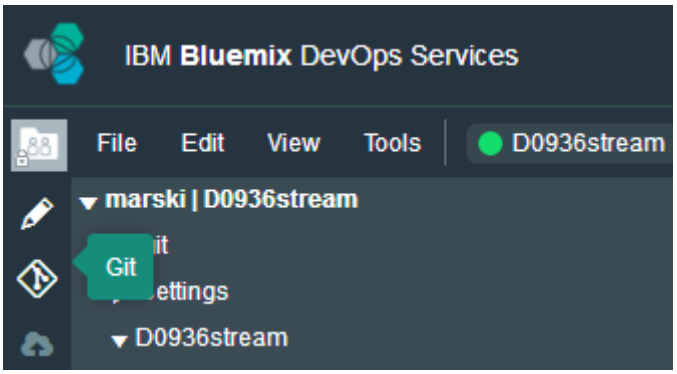
```
applications:
- disk_quota: 1024M
#  host: D0936stream
#  name: D0936stream
  path: webStarterApp.war
#  domain: stage1.mybluemix.net
  instances: 1
  memory: 512M
```

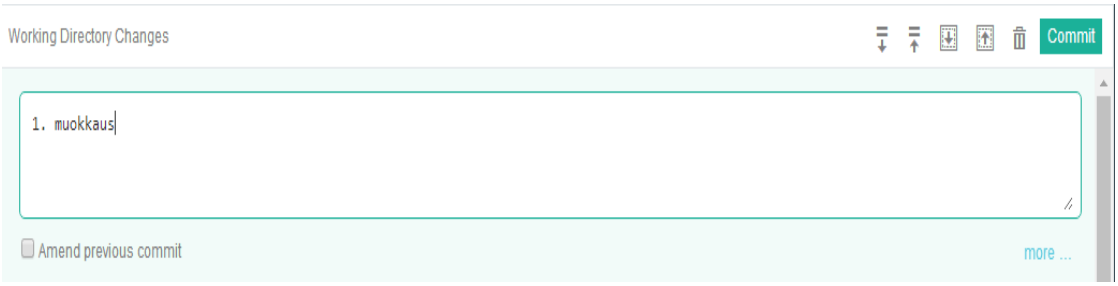
- Valitse sovelluksen tilarivin alasvetovalikosta ”Edit this launch configuration”. Ensimmäiseltä ”Manifest Settings” -sivulta yhdistä Streaming Analytics -palvelu ohjelmaan ja valitse ”Save”.

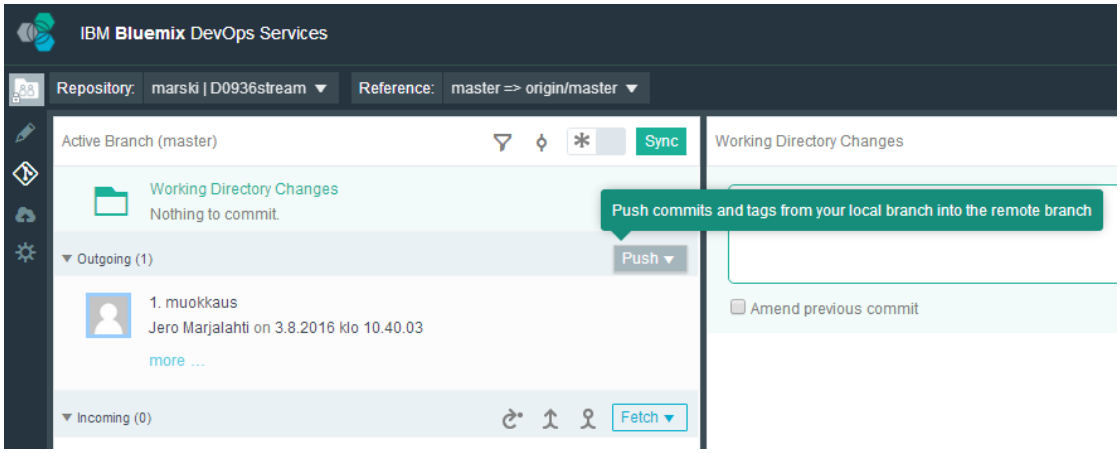




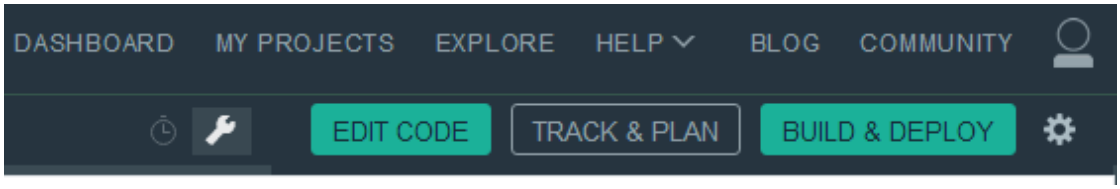
- Valitse vasemman reunan valikosta ”Git Repository”. Kirjoita edellisen kohdan muokkauksesta jokin viesti ja paina ”Commit”, ”Push” ja ”Fetch”.

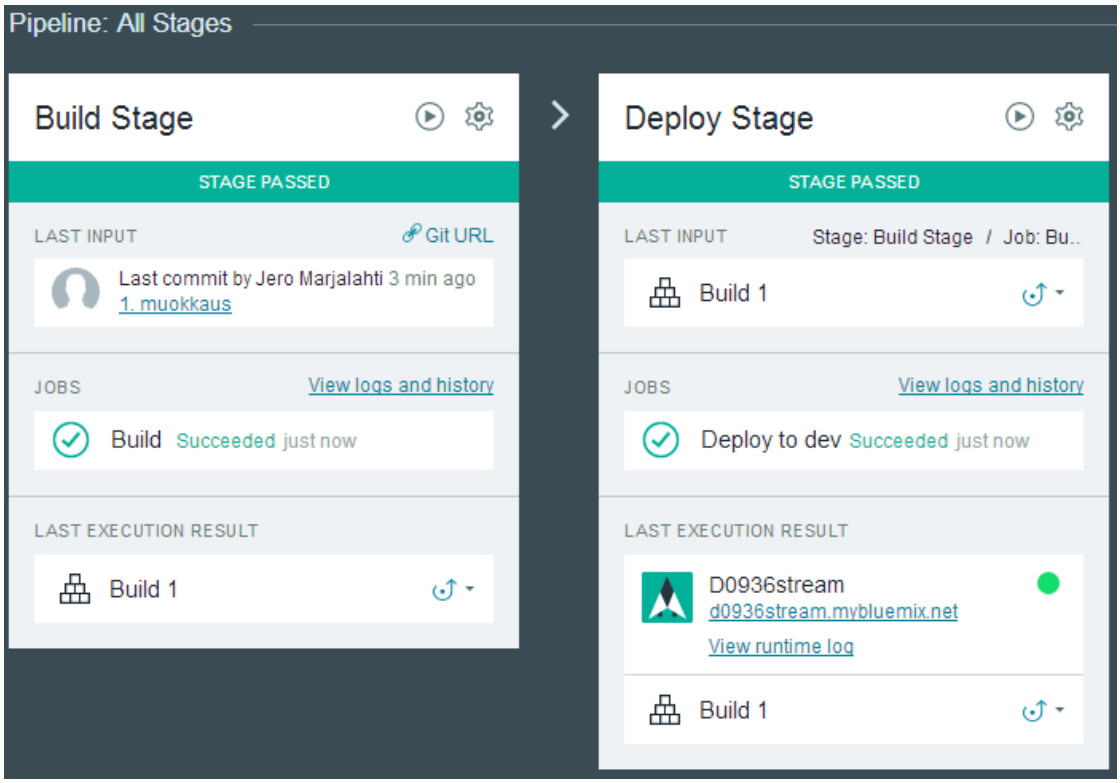






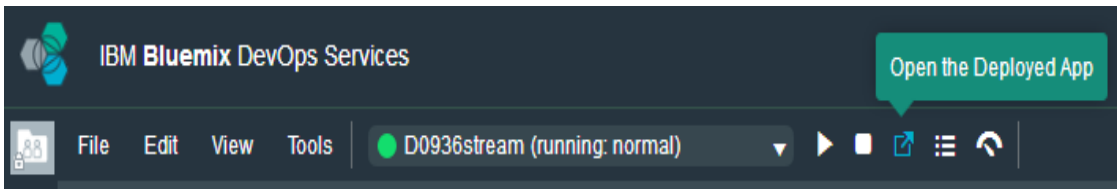
- Valitse ”Build and Deploy” oikealta ylhäältä ja anna ”Build Stage” ja ”Deploy Stage” -kohtien käydä muutokset läpi.





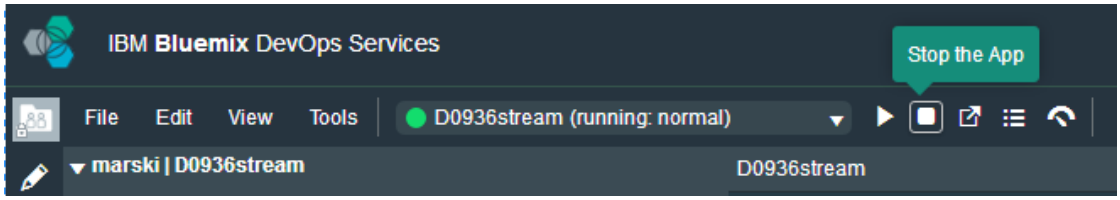
5. Sovelluksen ajaminen

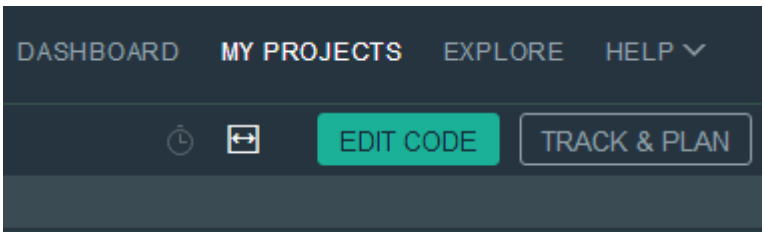
- Palaa DevOpsin etusivulle. Varmista, että sovelluksesi tila on ”running: normal”. Mikäli sovellus on pysähtynyt, valitse ”Deploy the App from the Workspace ”. Avaa sovelluksen verkkosivu valitsemalla ”Open the Deployed App”. New York Cityn liikennetietojen päivittyminen verkkosivulle voi kestää muutaman minuutin.

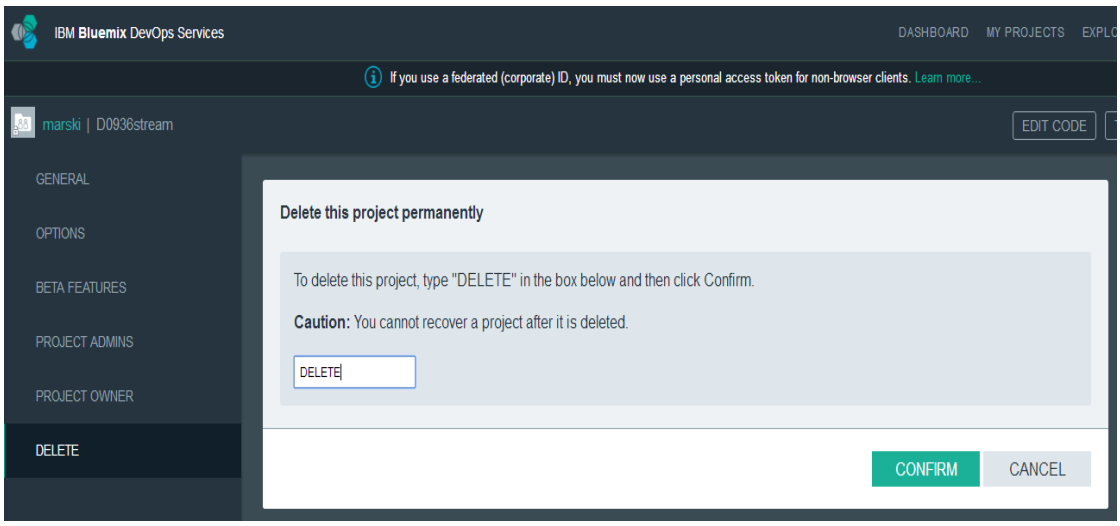


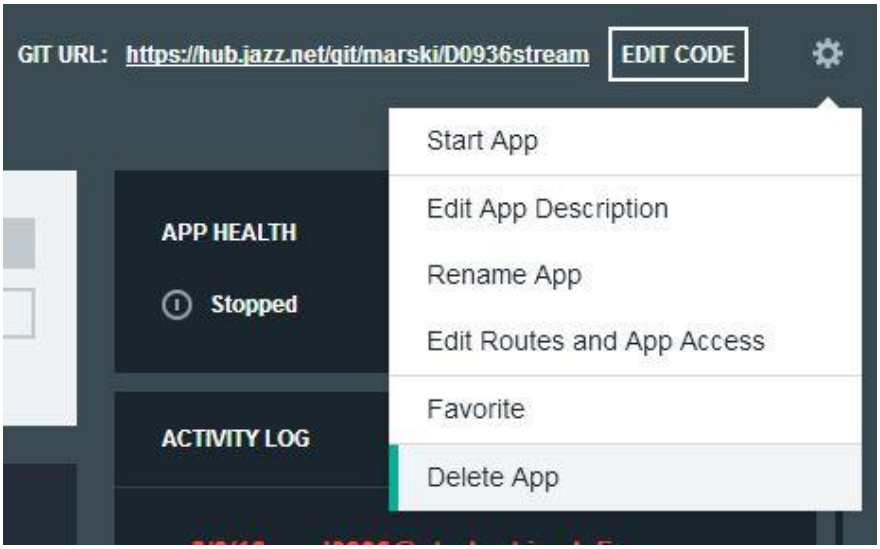
6. Sovelluksen pysäyttäminen ja poistaminen

- Sovellus pysäytetään valitsemalla ”Stop the App”. Projektin voi poistaa DevOpsista valitsemalla ”My Projects” ja valitsemalla kyseisen projektin vaihtoehdoista ”Delete”. Bluemixistä sovellus poistetaan ohjauspaneelin perusnäkymästä valitsemalla sovelluksen valikosta ”Delete App”.





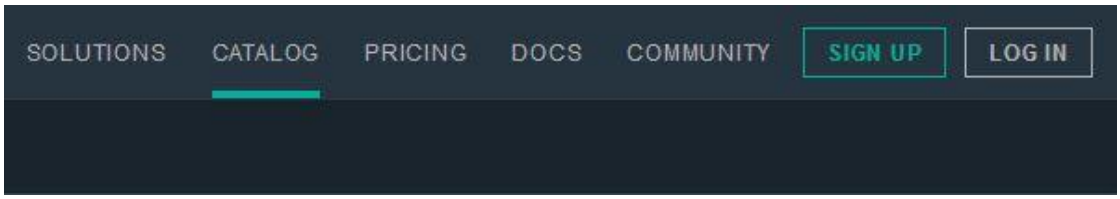



GIT URL: https://hub.jazz.net/git/marski/D0936stream
EDIT CODE
Start App
Edit App Description
Rename App
Edit Routes and App Access
Favorite
Delete App
APP HEALTH
Stopped
ACTIVITY LOG

Liite 2. Insights for Twitterin käyttöönotto

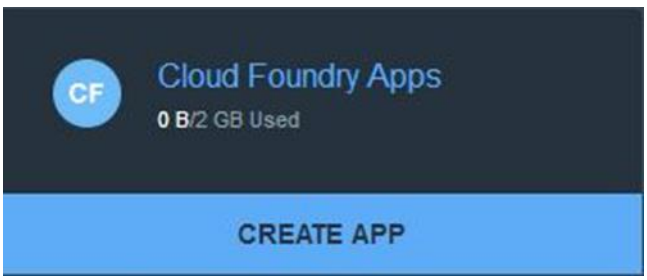
1. Tilien rekisteröinti

- Rekisteröi Bluemix-tili osoitteessa https://console.ng.bluemix.net/registration. Tämän jälkeen ota käyttöön DevOps-tili osoitteessa https://hub.jazz.net. Tilit ovat linkitetty yhteen henkilökohtaisen IBM ID:n kautta.

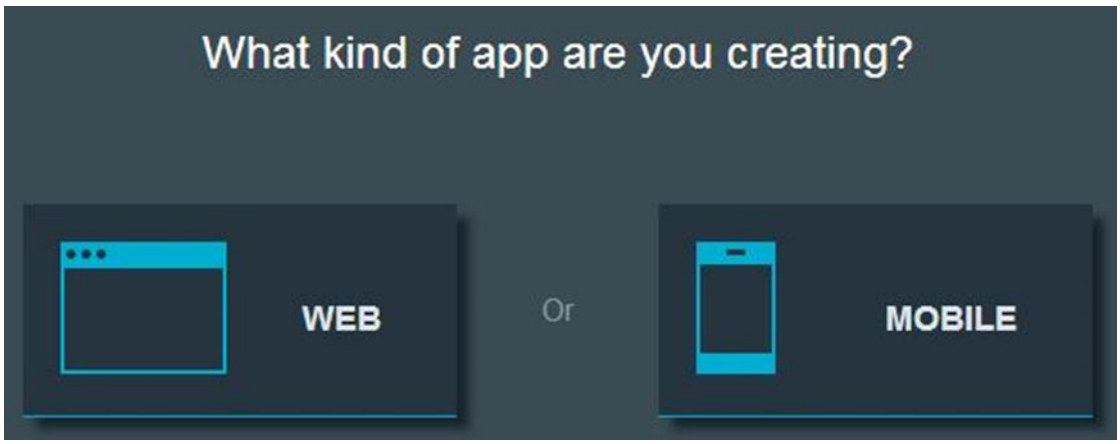


2. Bluemix sovelluksen luominen ja Insights for Twitterin yhdistäminen siihen

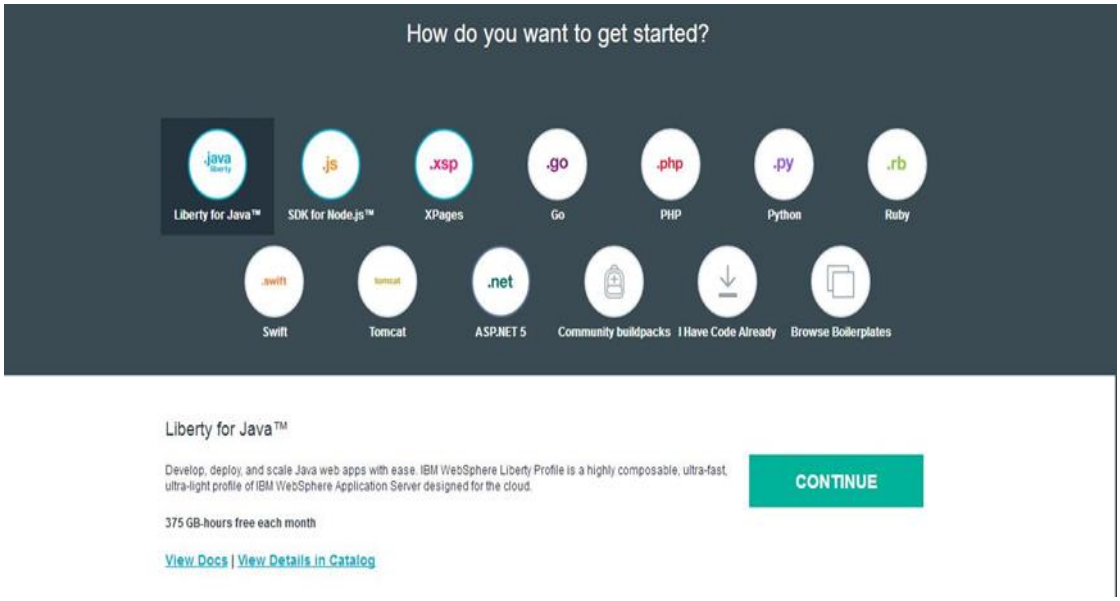
- Kirjaudu Bluemix-palveluun osoitteessa https://console.ng.bluemix.net ja lisää sovellus valitsemalla ”Create an App”.



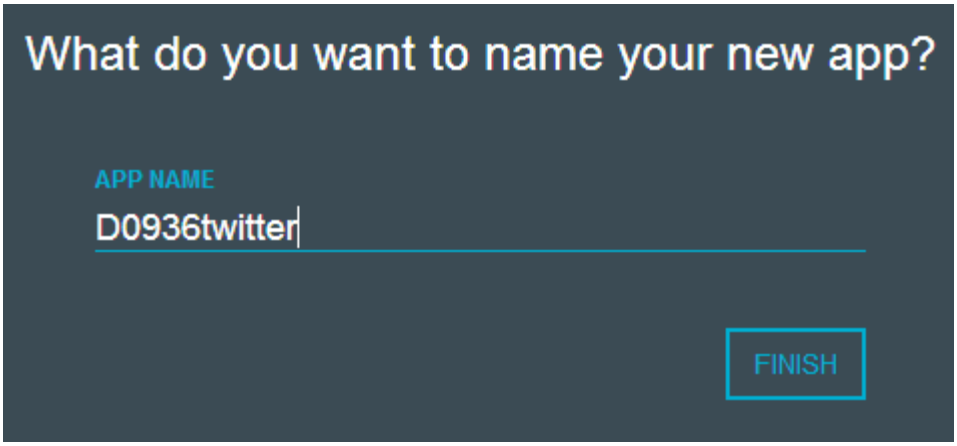
- Valitse sovelluksen tyypiksi ”Web”.



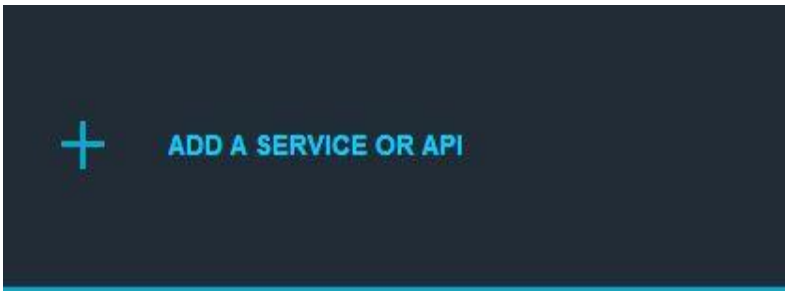
- Valitse käytettävissä olevista vaihtoehdoista ”Liberty for Java”.

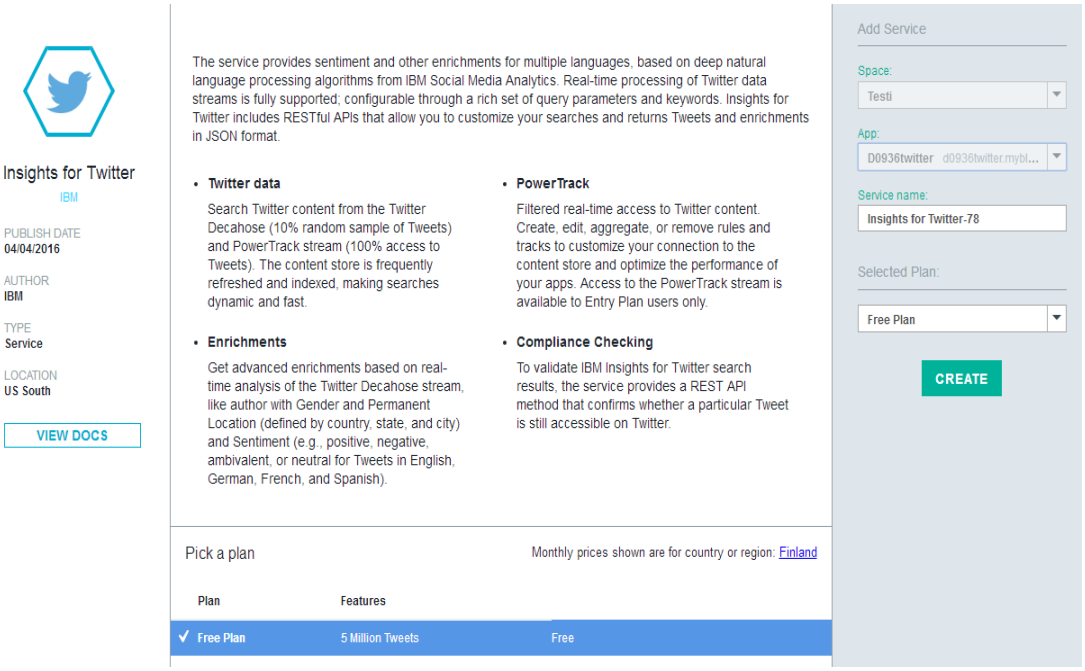


- Anna uudelle sovellukselle nimi.

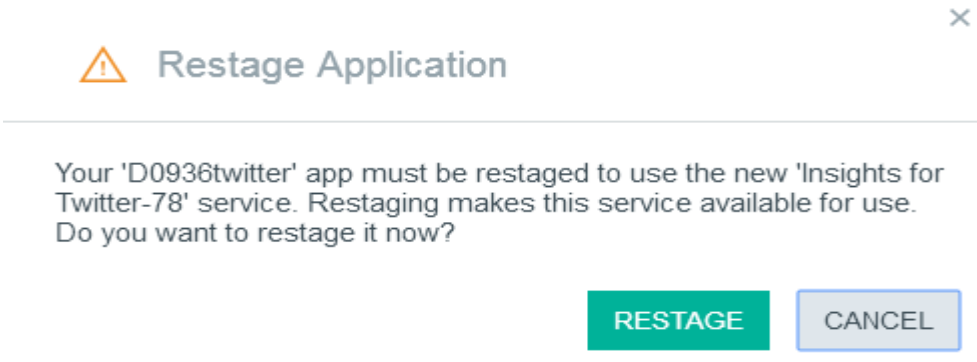


- Luodun ohjelman ”Overview” -sivulta valitse ”Add a Service or API”. Valitse palveluista Insights for Twitter ja valitse ”Create”.





- Valitse ”Restage” ilmoituksen tullessa näkyviin.



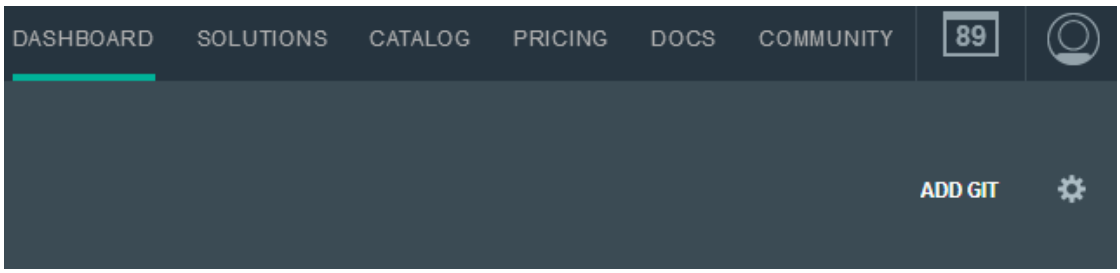
3. Lähdekoodin hankinta

- Lataa lähdekoodin .zip-tiedosto osoitteesta https://hub.jazz.net/project/kungr/sample-cdetestapp valitsemalla ”Download the contents of this branch as a zip file”.

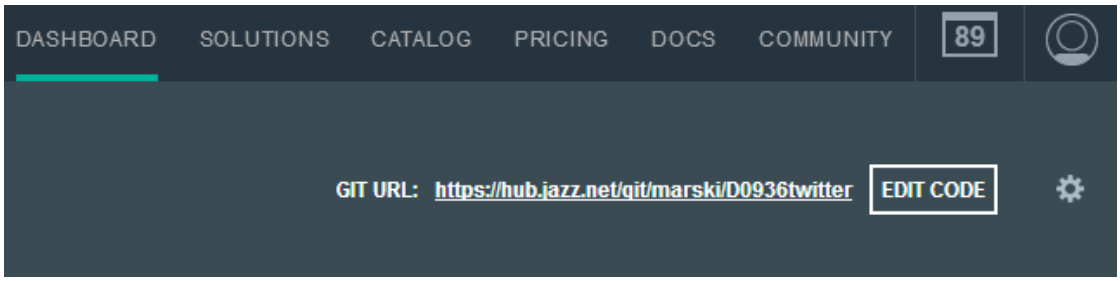
- Vaihtoehtoisesti voit myös ”forkata” alkuperäisen projektin tai kloonata git-tietolähteen. Alkuperäinen projekti löytyy DevOpsin valinnasta ”Explore” hakusanalla "sample-cdetestapp". Projektin nimi on “kungr | sample-cdetestapp”.

4. Sovelluksen käyttöönotto

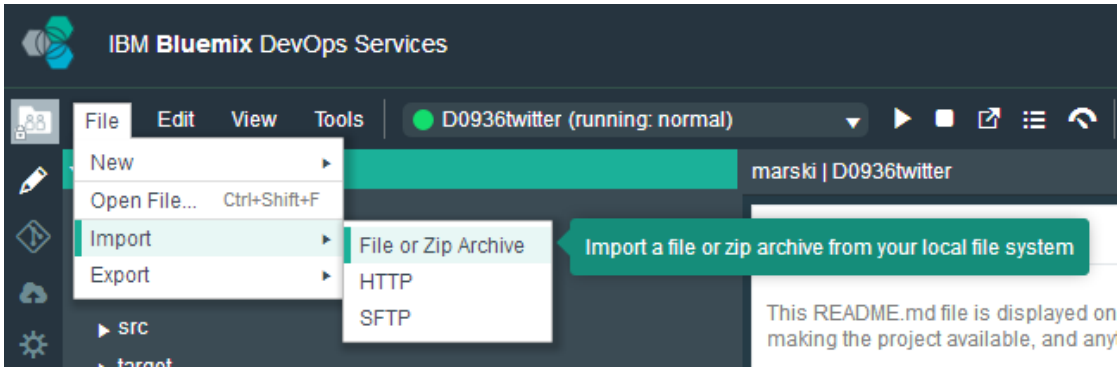
- Sovelluksen ”Overview” -sivulta valitse oikealta ylhäältä ”Add Git” ja valitse ”Continue”.



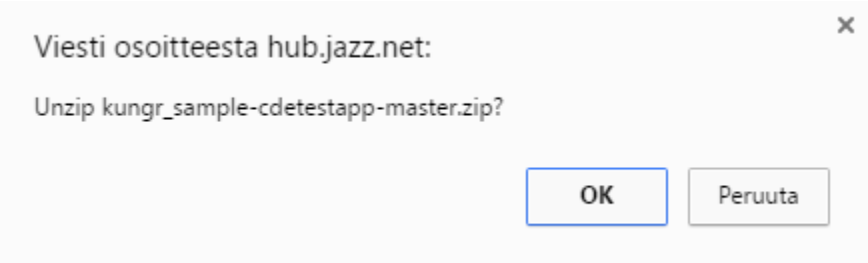
- Sovelluksen ”Overview” -sivulta valitse oikealta ylhäältä ”Edit Code”.



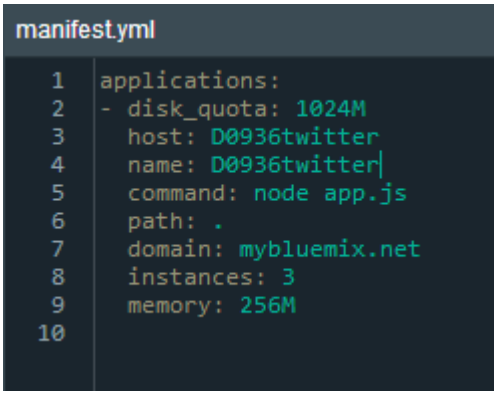
- Tuo lataamasi lähdekoodi projektiin valitsemalla ”File - Import - File or ZIP Archive”.



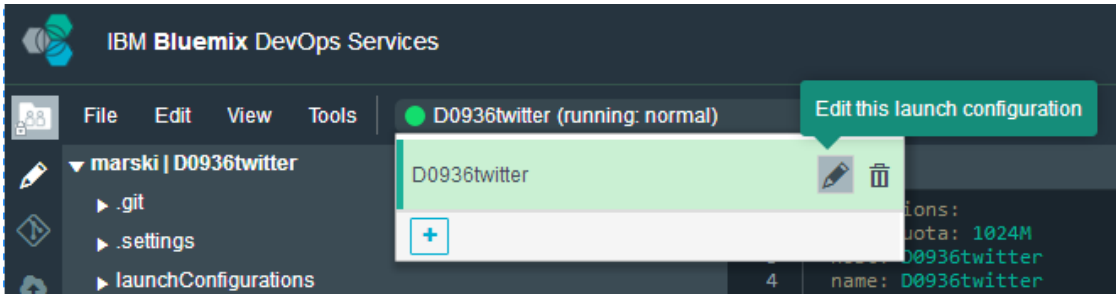
- Anna DevOpsin purkaa .zip-tiedosto latauksen yhteydessä.

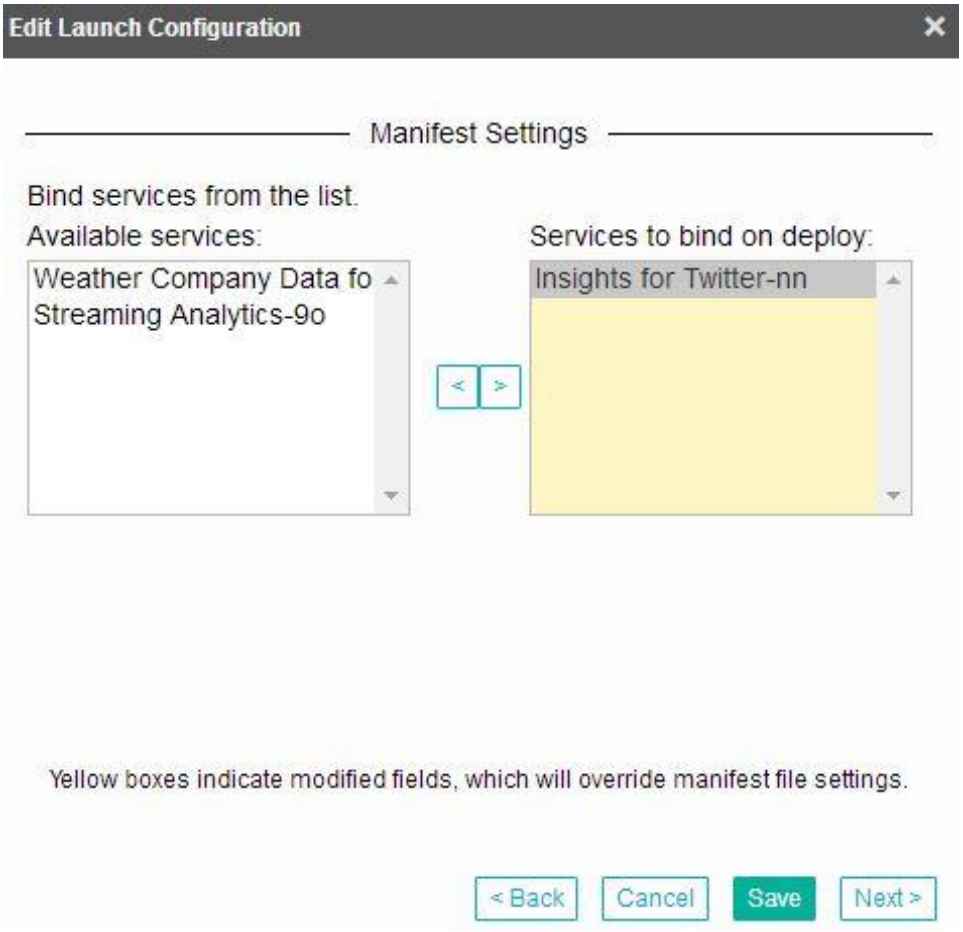


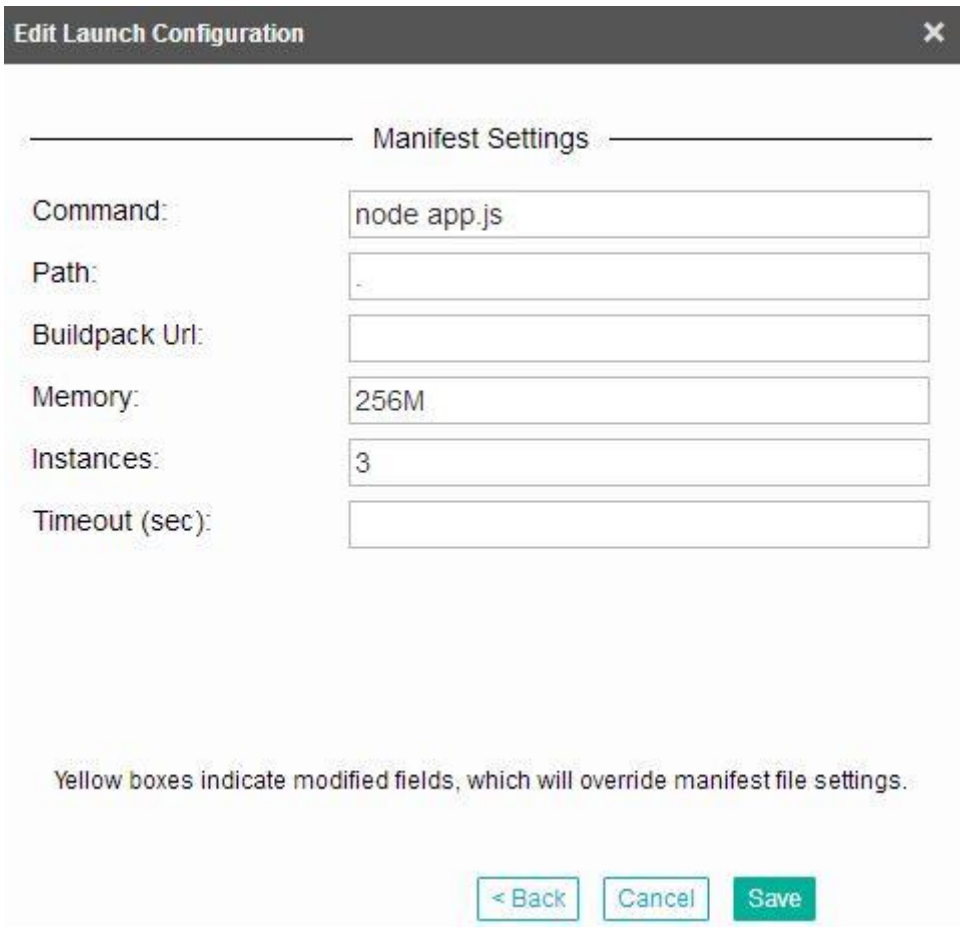
- Muokkaa manifest.yml tiedostoon host- ja name-riveille sovelluksesi nimi.



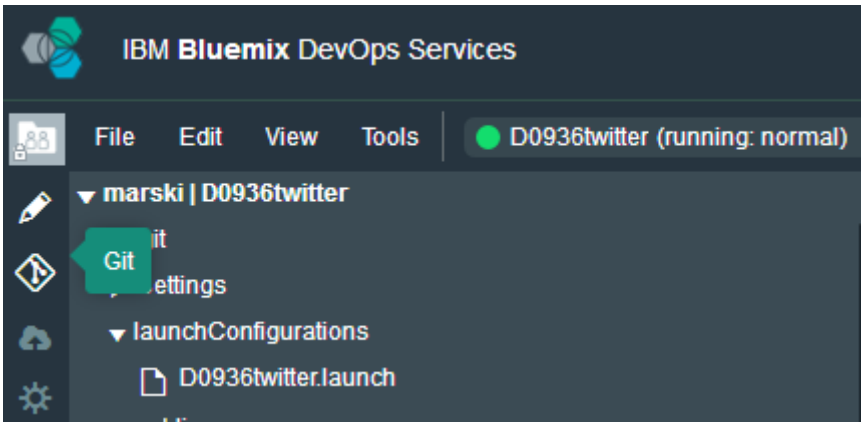
- Valitse sovelluksen tilarivin alasvetovalikosta ”Edit this launch configuration”. Ensimmäiseltä ”Manifest Settings” -sivulta yhdistä Insights for Twitter -palvelu ohjelmaan. Toiselta ”Manifest Settings” -sivulta muokkaa path-riville ”.” ja valitse ”Save”.

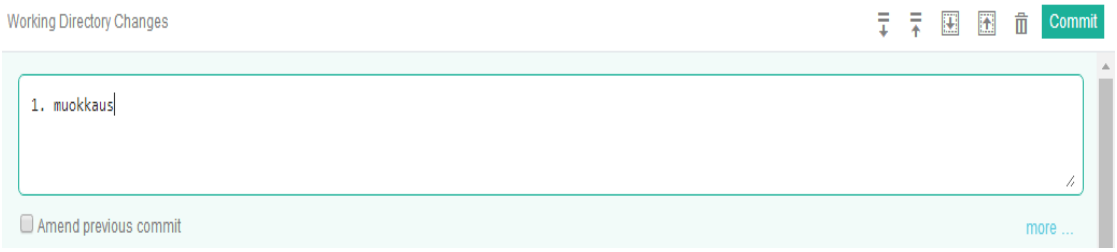


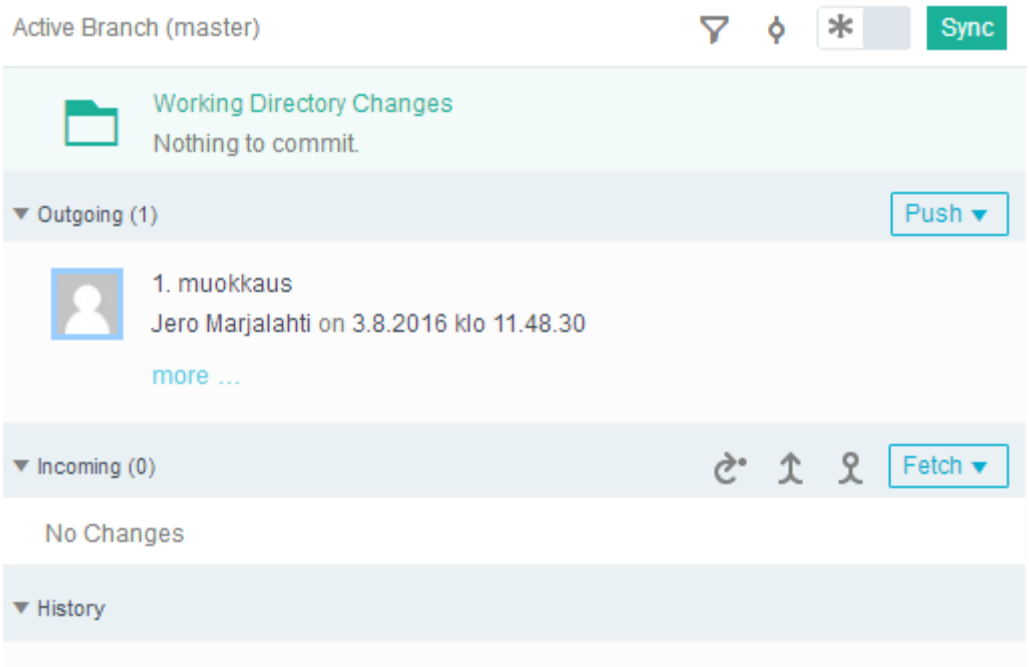

Edit Launch Configuration
Manifest Settings
Bind services from the list.
Available services:
Weather Company Data fo
Streaming Analytics-9o
Services to bind on deploy:
Insights for Twitter-nn
< >
Yellow boxes indicate modified fields, which will override manifest file settings.
< Back
Cancel
Save
Next >


Edit Launch Configuration
Manifest Settings
Command:
node app.js
Path:
.
Buildpack Url:
Memory:
256M
Instances:
3
Timeout (sec):
Yellow boxes indicate modified fields, which will override manifest file settings.
< Back
Cancel
Save

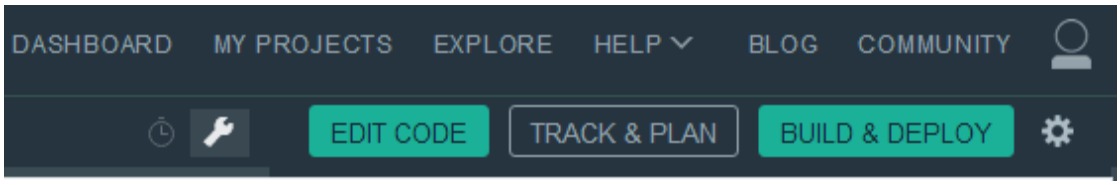
- Valitse vasemman reunan valikosta ”Git Repository”. Kirjoita edellisen kohdan muokkauksesta jokin viesti ja paina ”Commit”, ”Push” ja ”Fetch”.

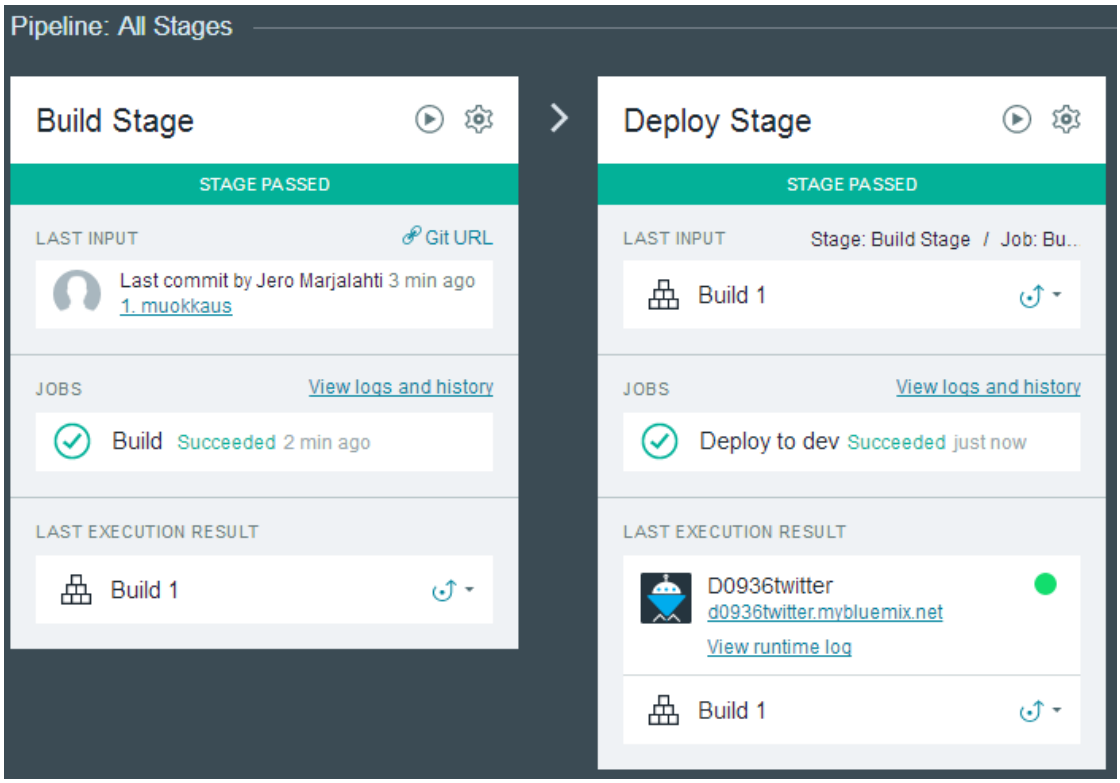






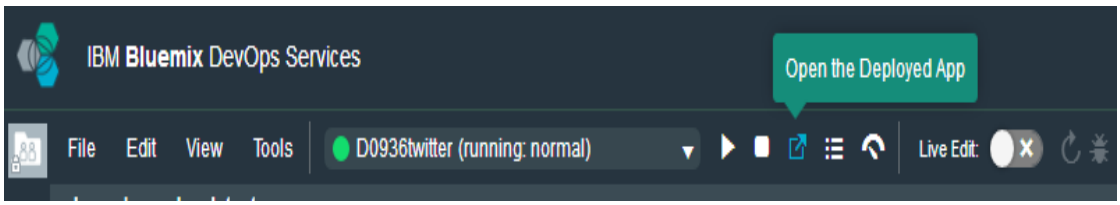
- Valitse ”Build and Deploy” oikealta ylhäältä ja anna ”Build Stage” ja ”Deploy Stage” -kohtien käydä muutokset läpi.





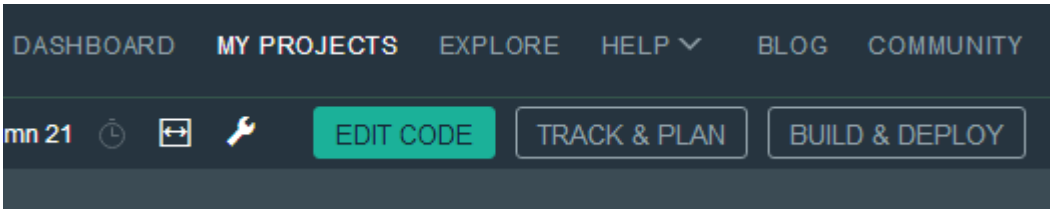
5. Sovelluksen ajaminen

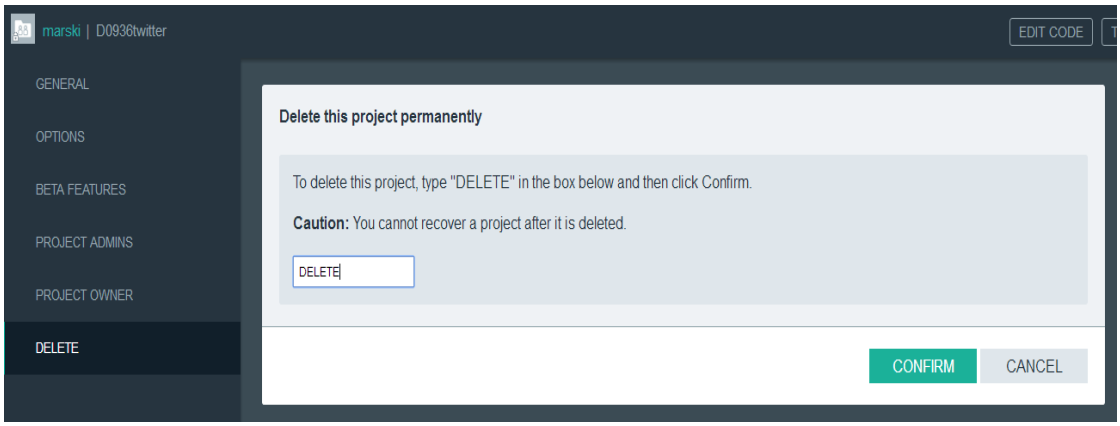
- Palaa DevOpsin etusivulle. Varmista, että sovelluksesi tila on ”running: normal”. Mikäli sovellus on pysähtynyt, valitse ”Deploy the App from the Workspace ”. Avaa Twitter Decahosea käyttävän hakukoneen verkkosivu valitsemalla ”Open the Deployed App”.

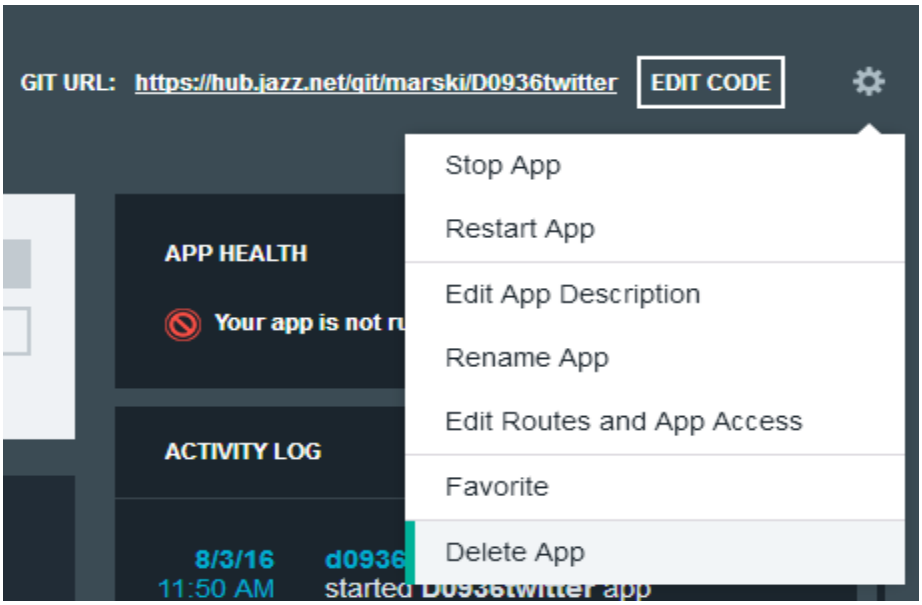


6. Sovelluksen pysäyttäminen ja poistaminen

- Sovellus pysäytetään valitsemalla ”Stop the App”. Projektin voi poistaa DevOpsista valitsemalla ”My Projects” ja valitsemalla kyseisen projektin vaihtoehdoista ”Delete”. Bluemixistä sovellus poistetaan ohjauspaneelin perusnäkymästä valitsemalla sovelluksen valikosta ”Delete App”.



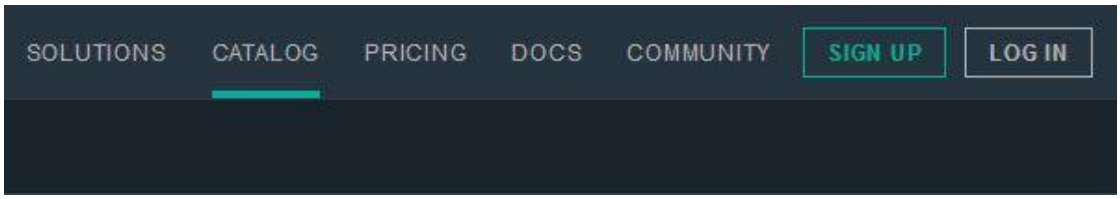




Liite 3. Weather Company Data for IBM Bluemixin käyttöönotto

1. Tilien rekisteröinti

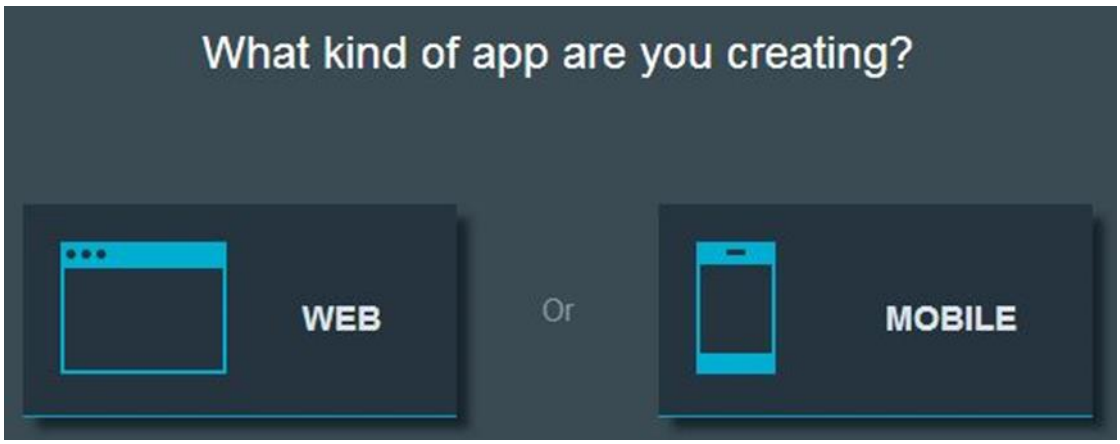
- Rekisteröi Bluemix-tili osoitteessa https://console.ng.bluemix.net/registration. Tämän jälkeen ota käyttöön DevOps-tili osoitteessa https://hub.jazz.net. Tilit ovat linkitetty yhteen henkilökohtaisen IBM ID:n kautta.



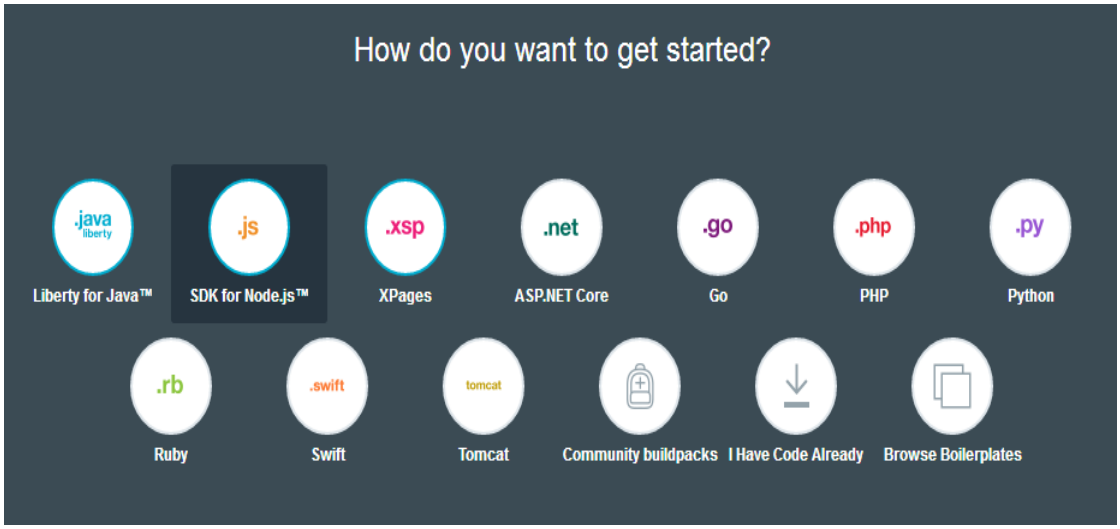
2. Bluemix sovelluksen luominen ja Weather Company Data for IBM Bluemixin yhdistäminen siihen

- Kirjaudu Bluemix-palveluun osoitteessa https://console.ng.bluemix.net ja lisää sovellus valitsemalla ”Create an App”.

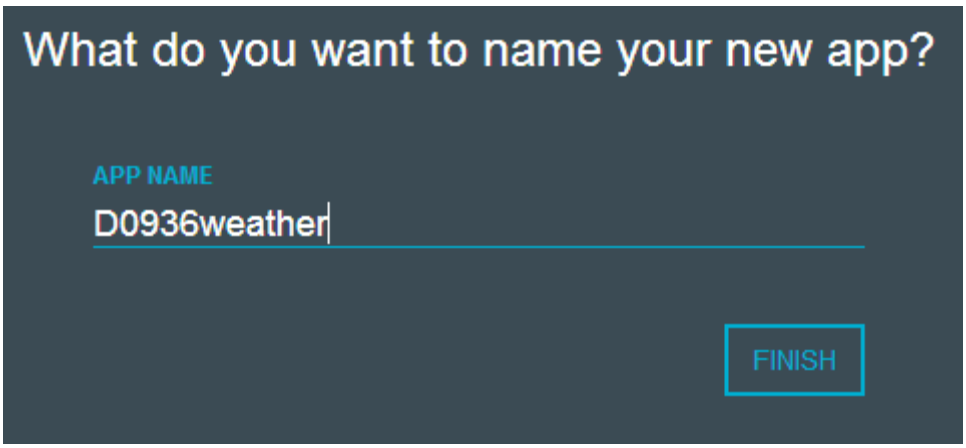
- Valitse sovelluksen tyypiksi ”Web”.



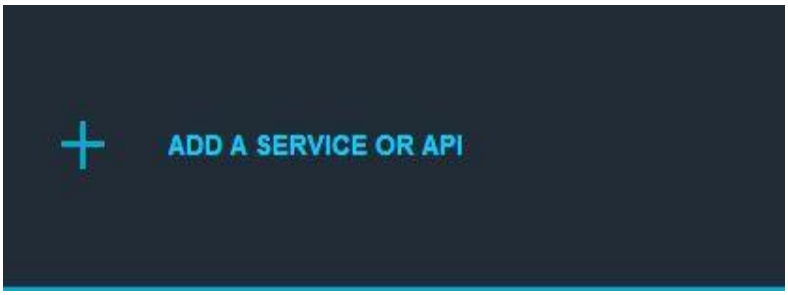
- Valitse käytettävissä olevista vaihtoehdoista ”SDK for Node.js”.

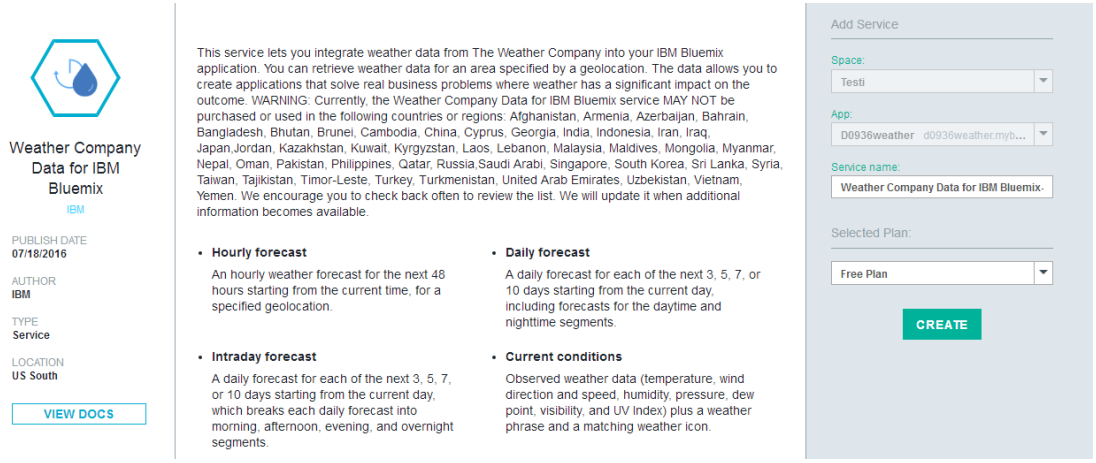


- Anna uudelle sovellukselle nimi.

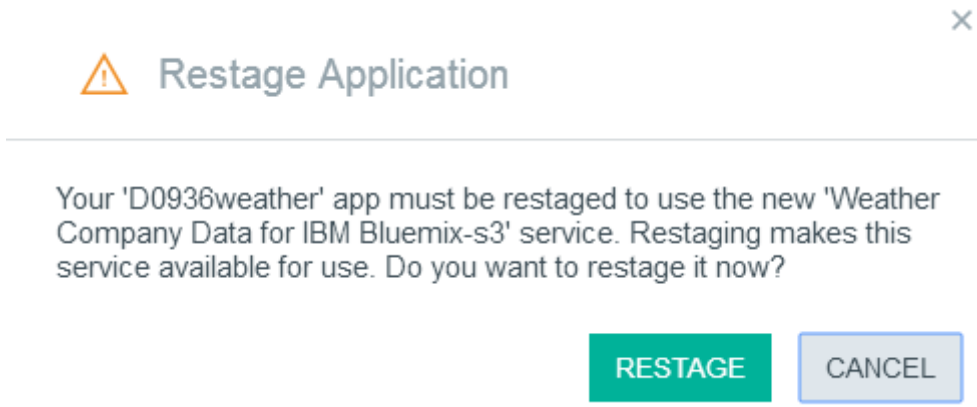


- Luodun ohjelman ”Overview” -sivulta valitse ”Add a Service or API”. Valitse palveluista Weather Company Data for IBM Bluemix ja valitse ”Create”.





- Valitse ”Restage” ilmoituksen tullessa näkyviin.

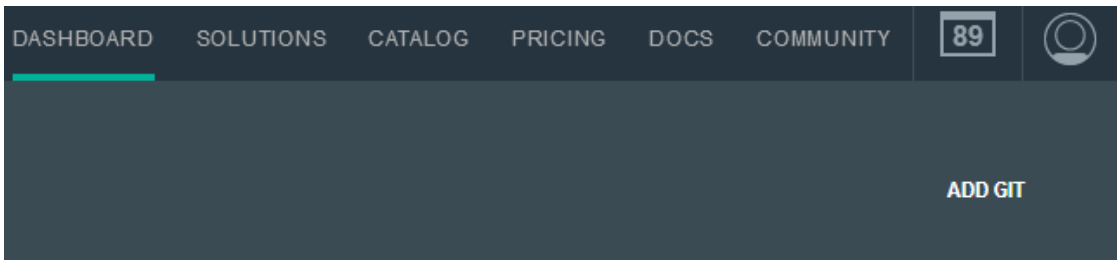


3. Lähdekoodin hankinta

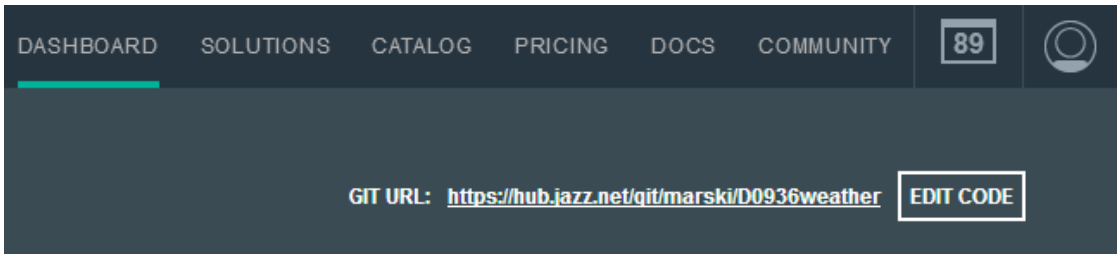
- Lataa lähdekoodin .zip-tiedosto osoitteesta https://github.com/IBM-Bluemix/weather-company-data-demo valitsemalla ”Clone or Download - Download ZIP”. Vaihtoehtoisesti voit myös kloonata sovelluksen samasta valikosta tai ottaa sovelluksen suoraan käyttöön valitsemalla ”Deploy to Bluemix”.

4. Sovelluksen käyttöönotto

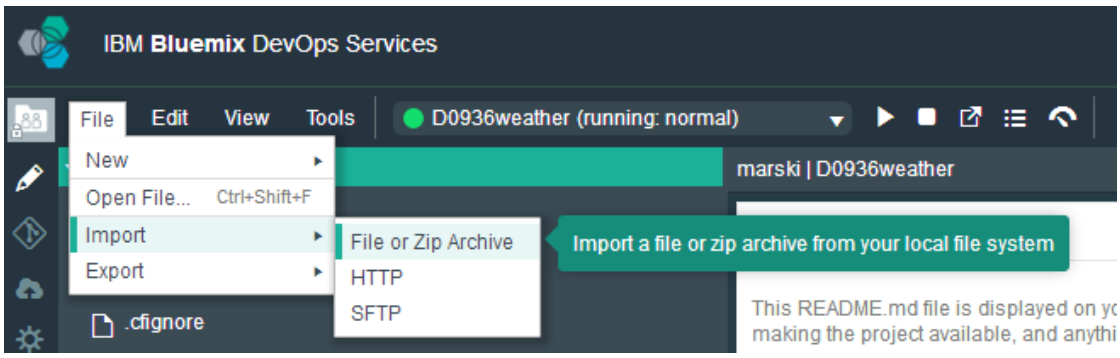
- Sovelluksen ”Overview” -sivulta valitse oikealta ylhäältä ”Add Git” ja valitse ”Continue”.



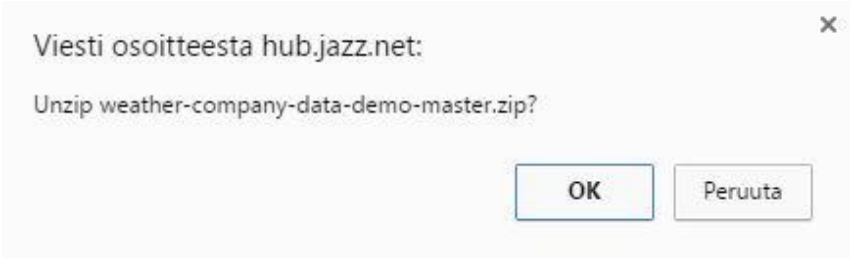
- Sovelluksen ”Overview” -sivulta valitse oikealta ylhäältä ”Edit Code”.

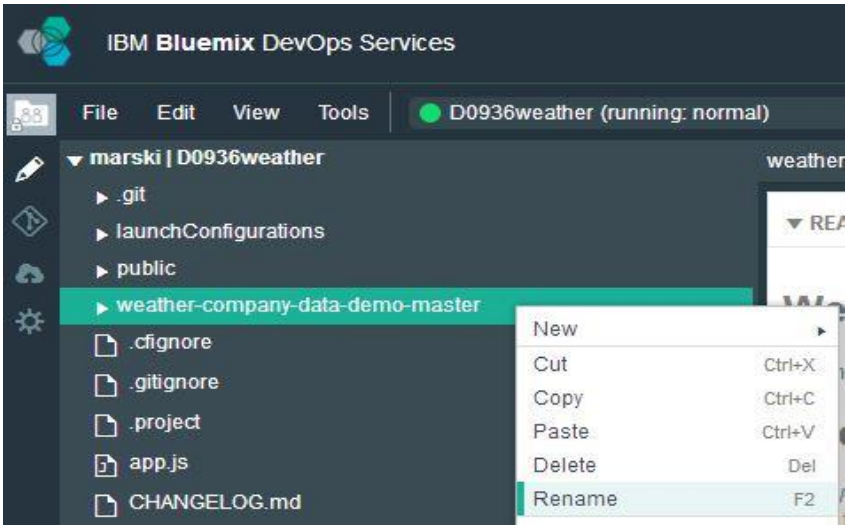


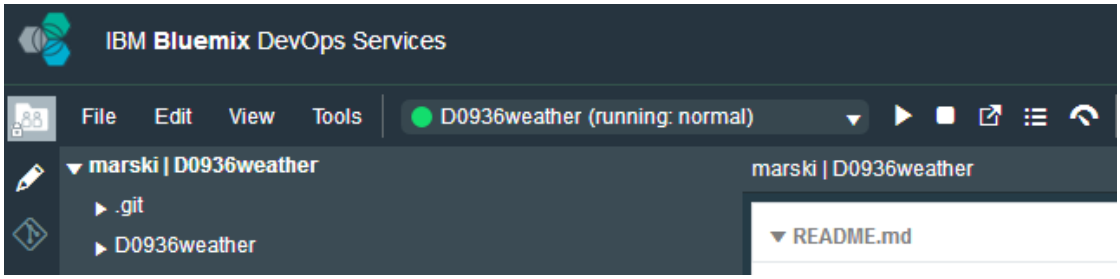
- Tuo lataamasi lähdekoodi projektiin valitsemalla ”File - Import - File or ZIP Archive”.



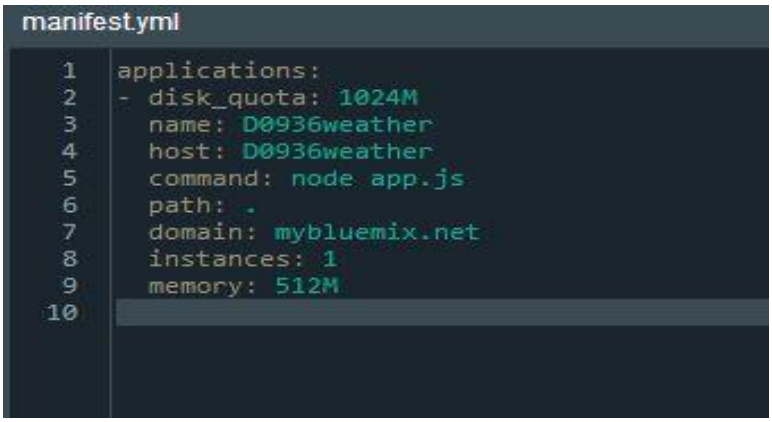
- Anna DevOpsin purkaa .zip-tiedosto latauksen yhteydessä ja nimeä purettu kansio ”weather-company-data-demo-master” kohdassa kaksi annetun uuden sovelluksen nimen mukaisesti.



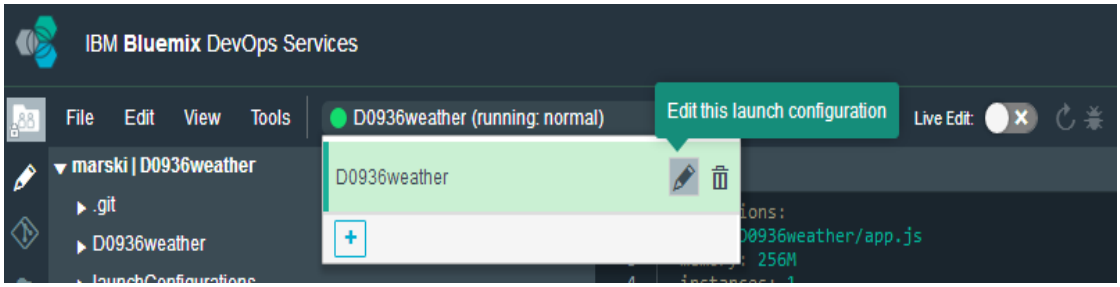


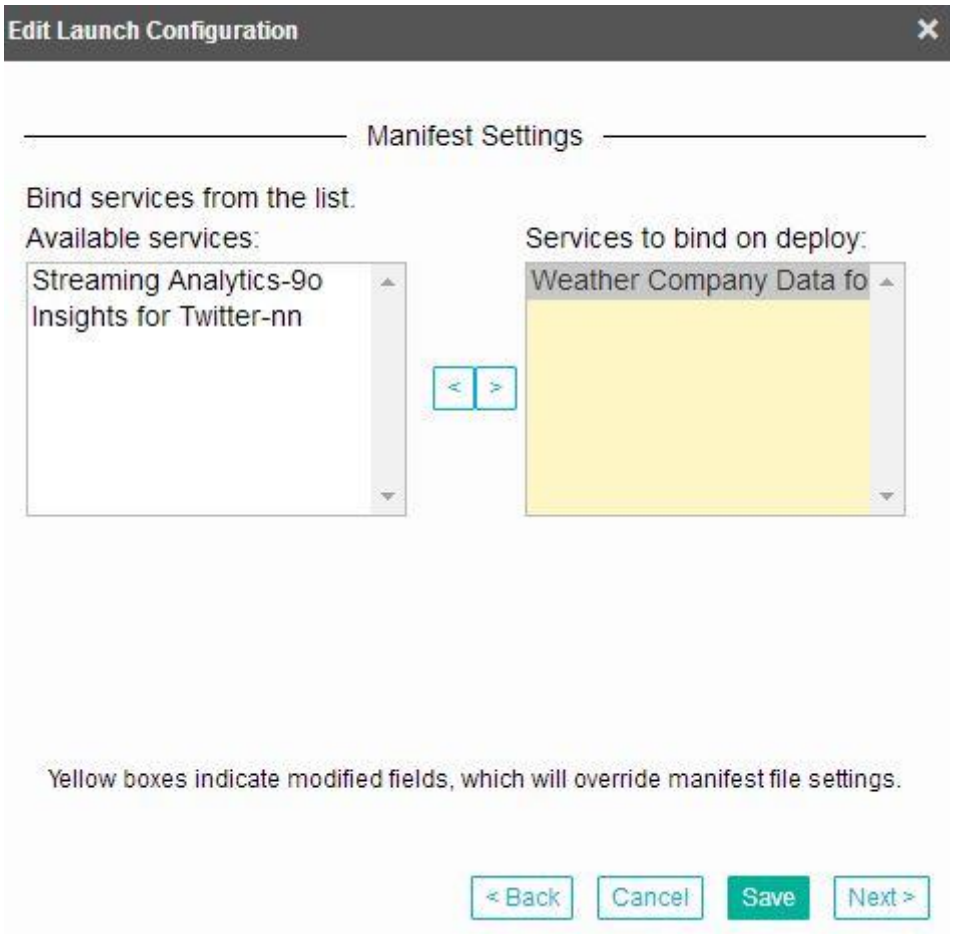


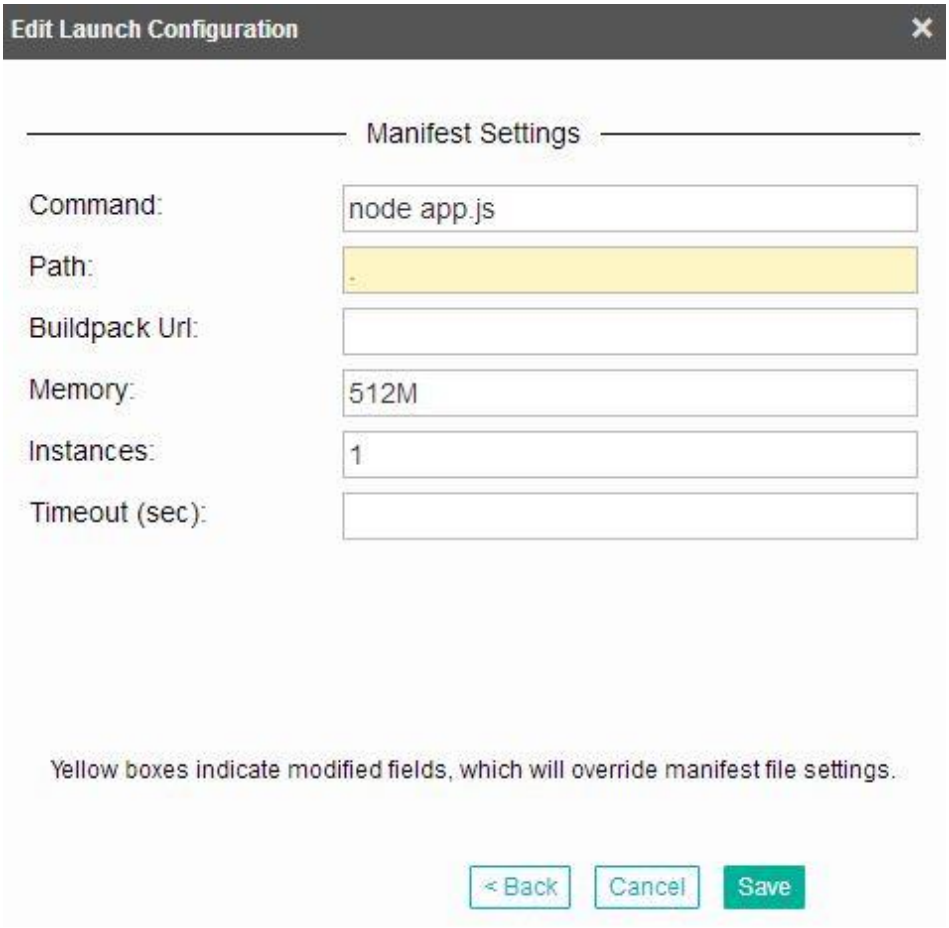
- Muokkaa puretun kansion nimi/manifest.yml tiedostoon host- ja name-riveille oman sovelluksesi nimi.



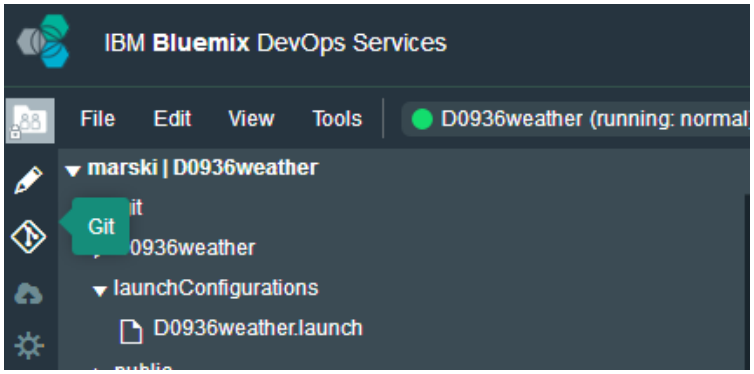
- Valitse sovelluksen tilarivin alasvetovalikosta ”Edit this launch configuration”. ”Edit Launch Configuration” -sivulta muuta Manifest File -kohtaan puretun kansion nimi/manifest.yml. Ensimmäiseltä ”Manifest Settings” sivulta yhdistä Weather Company Data for IBM Bluemix -palvelu ohjelmaan. Älä tee muutoksia toiselle ”Manifest Settings”-sivulle vaan valitse ”Save”.

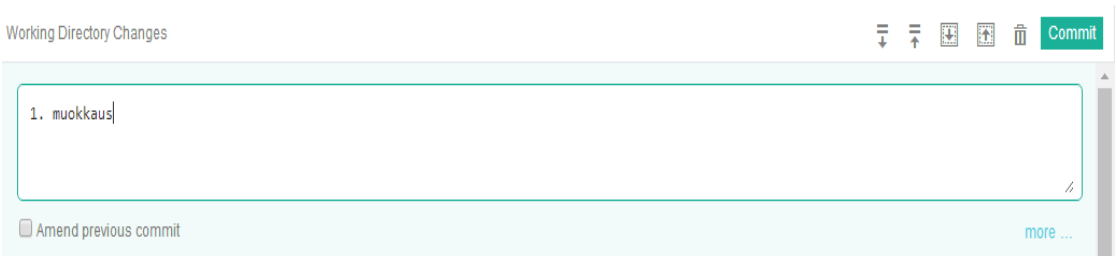


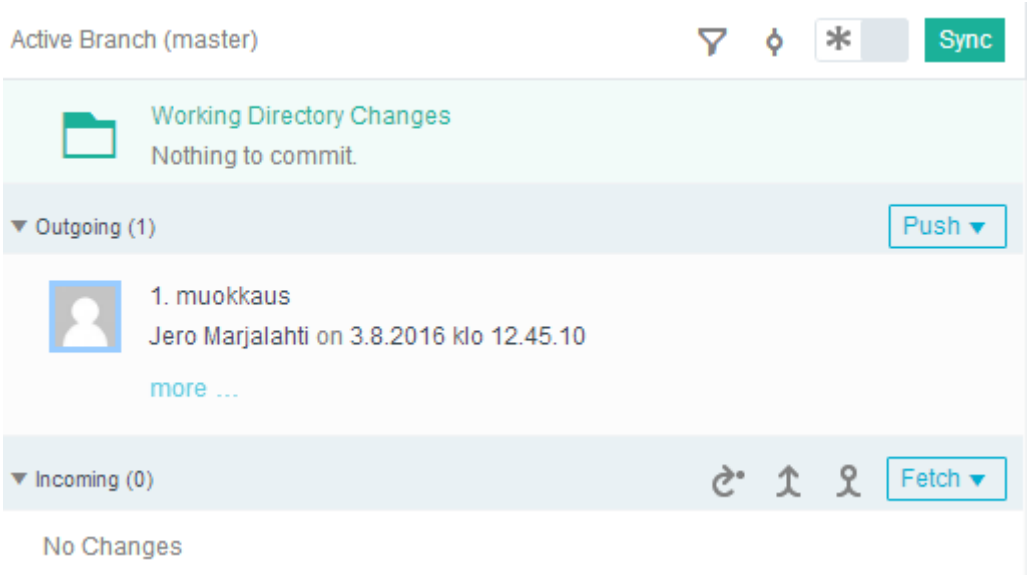




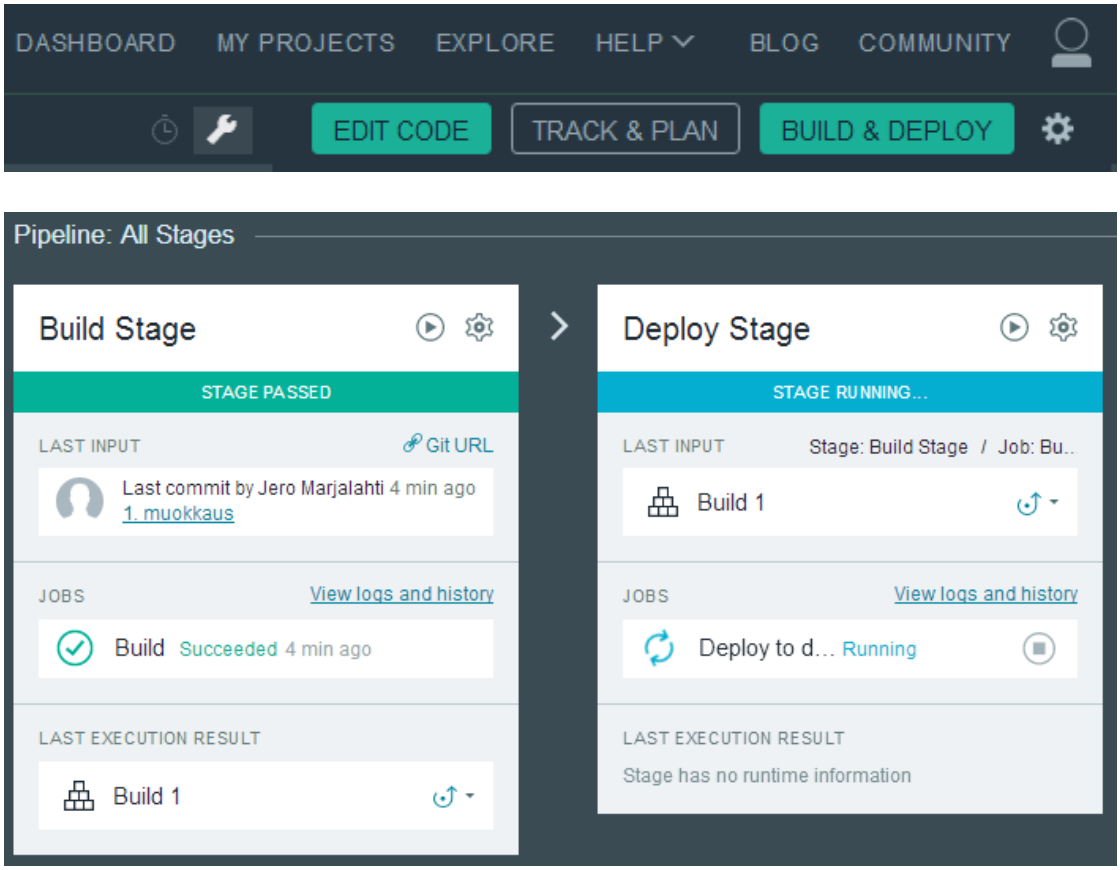
- Valitse vasemman reunan valikosta ”Git Repository”. Kirjoita edellisen kohdan muokkauksesta jokin viesti ja paina ”Commit”, ”Push” ja ”Fetch”.





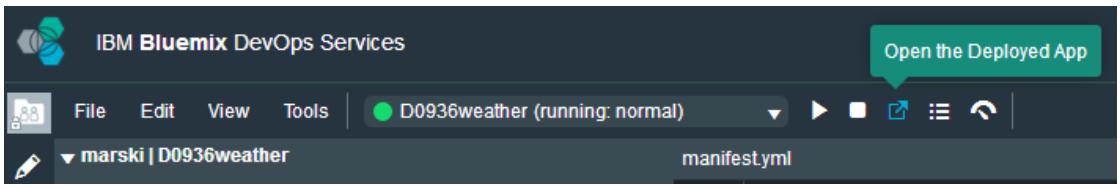


- Valitse ”Build and Deploy” oikealta ylhäältä ja anna ”Build Stage” ja ”Deploy Stage” kohtien käydä muutokset läpi.



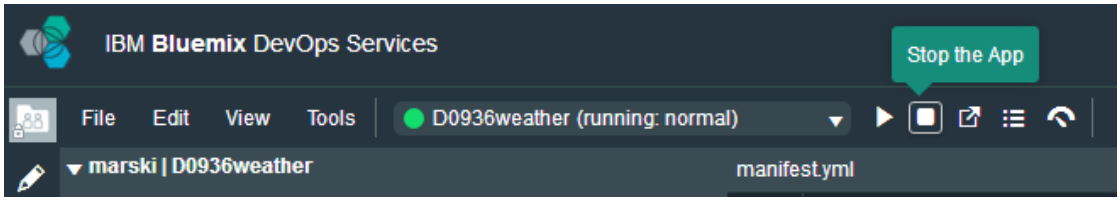
5. Sovelluksen ajaminen

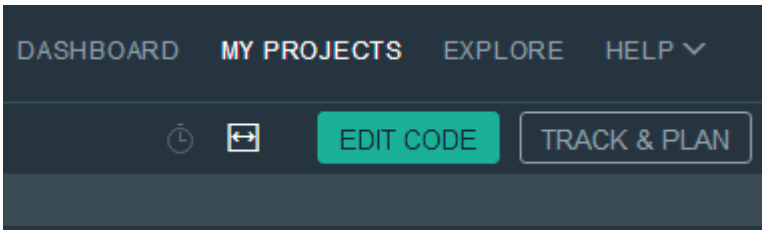
- Palaa DevOpsin etusivulle. Varmista, että sovelluksesi tila on ”running: normal”. Mikäli sovellus on pysähtynyt, valitse ”Deploy the App from the Workspace ”. Avaa The Weather Companyn dataa käyttävän sääsovelluksen verkkosivu valitsemalla ”Open the Deployed App”

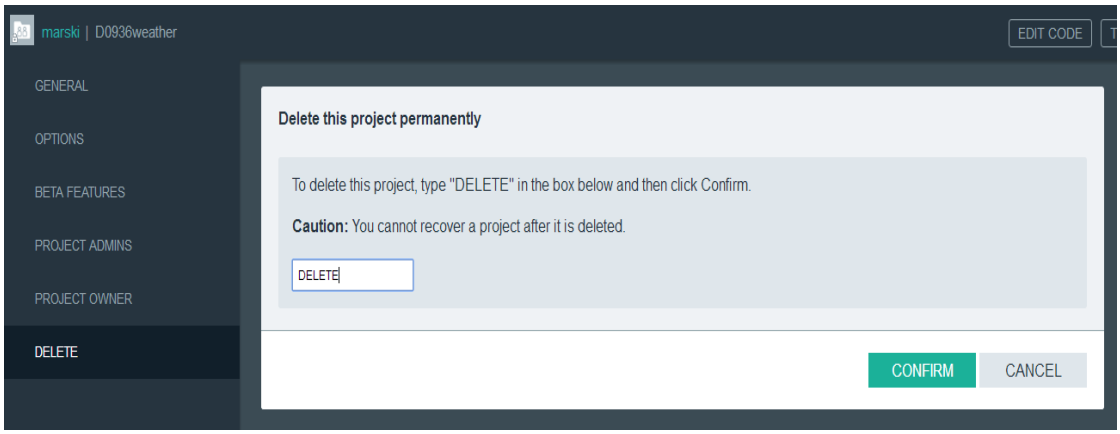


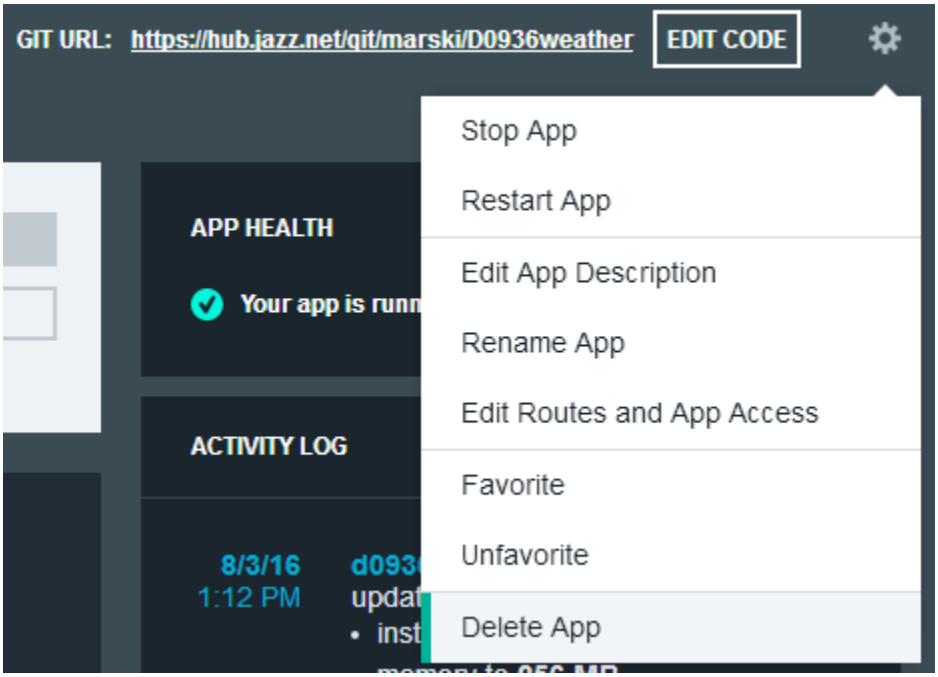
6. Sovelluksen pysäyttäminen ja poistaminen

- Sovellus pysäytetään valitsemalla ”Stop the App”. Projektin voi poistaa DevOpsista valitsemalla ”My Projects” ja valitsemalla kyseisen projektin vaihtoehdoista ”Delete”. Bluemixistä sovellus poistetaan ohjauspaneelin perusnäkymästä valitsemalla sovelluksen valikosta ”Delete App”.








GIT URL: https://hub.jazz.net/git/marski/D0936weather
EDIT CODE
Stop App
Restart App
Edit App Description
Rename App
Edit Routes and App Access
Favorite
Unfavorite
Delete App
APP HEALTH
Your app is runn
ACTIVITY LOG
8/3/16
1:12 PM